Spatial sampling in ensemble perception of hue

Lari Sakari Virtanen

Master's Thesis

Psychology

Faculty of Medicine

January 2019

Supervisor: Toni Saarela

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department
Faculty of Medicine / Department of Psychology and Logopedics	

Tekijä/Författare – Author
Lari Virtanen

Työn nimi / Arbetets titel – Title
Spatial sampling in ensemble perception of hue

Oppiaine /Läroämne – Subject
Psychology

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Master's Thesis	January 2019	43

Tiivistelmä/Referat – Abstract

An object's distinct color is commonly formed from a chromatic distribution, with varying hue information. Human observers can estimate the mean hue of a hue ensemble, but the details of this ability are still relatively unknown. Ensemble perception of hue was studied by systematically varying the external noise and the amount of information available. In a set of four experiments, the spatial sampling characteristics, spontaneous estimation strategy and task dependence, were examined.

Presented briefly against a gray background, the stimuli consisted of a number of square elements, each with a uniform hue. The hues were drawn either from a von Mises distribution or a highly skewed normal distribution with one of three levels of external noise, located on a hue circle in CIELAB color space. In Experiments 1-3 observers performed a two-interval forced choice task and in Experiment 4 a modified match-to-sample task. Discrimination thresholds were estimated by fitting psychometric functions to the data. The number of elements utilized in averaging was estimated through equivalent noise modeling.

Discrimination thresholds increased with increasing external noise, but decreased as the number of elements increased, the improvement being greater with higher noise. Modeling the number of samples used by the observer as a fixed power of the samples available gave an excellent fit to the data. For the 64-element stimulus, estimated effective number of samples ranged from 16 to 41. Control experiments confirmed that performance improved with the number of elements, not stimulus area. Even with a highly skewed hue distribution, most observers opted for a simple averaging strategy in estimation. With a less straightforward comparison task, results appeared much less consistent.

Observers sample and average stimulus elements to estimate mean hue, similarly to ensemble perception in other domains. The observed sampling clearly surpasses earlier estimates pointing towards global sampling, and performance is most affected by surface, not edge information. Simple averaging remained as the observers' spontaneous strategy with a non-normal hue distribution. The results were also dependent on task details, emphasizing caution when comparing different kinds of experiments.

Avainsanat – Nyckelord – Keywords
hue averaging, ensemble perception, color vision

Säilytyspaikka – Förvaringställe – Where deposited
Helsinki University Library – Helda / E-thesis (theses) *https://ethesis.helsinki.fi/*

Muita tietoja – Övriga uppgifter – Additional information

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiedekunta/Osasto Fakultet/Sektion – Faculty	Laitos/Institution– Department
Lääketieteellinen tiedekunta / psykologian ja logopedian osasto	

Tekijä/Författare – Author
Lari Virtanen

Työn nimi / Arbetets titel – Title
Spatiaalinen otanta värisävyn joukkohavainnossa

Oppiaine /Läroämne – Subject
Psykologia

Työn laji/Arbetets art – Level	Aika/Datum – Month and year	Sivumäärä/ Sidoantal – Number of pages
Pro gradu -tutkielma	Tammikuu 2019	43

Tiivistelmä/Referat – Abstract

Objektin ominainen väri muodostuu yleisesti värijakaumasta, vaihtelevalla sävy-informaatiolla. Havainnoijat pystyvät arvioimaan värisävyjoukon keskimääräisen sävyn, mutta tämän kyvyn yksityiskohdat ovat edelleen verrattain tuntemattomia. Värisävyn joukkohavainnointia tutkittiin varioimalla systemaattisesti ulkoista kohinaa ja informaation määrää. Neljän kokeen sarjassa tarkasteltiin spatiaalisen otannan piirteitä, spontaania estimointistrategiaa sekä tehtäväriippuvuutta.

Esitettynä lyhyesti harmaata taustaa vasten, ärsykkeet koostuivat joukosta neliönmuotoisia, tasasävyisiä elementtejä. Värisävyt poimittiin joko von Mises -jakaumasta tai voimakkaasti vinosta normaalijakaumasta yhdellä kolmesta kohinatasosta, sijoitettuna sävy-ympyrälle CIELAB väriavaruudessa. Kokeissa 1-3 havainnoijat suorittivat kahden intervallin pakkovalintatehtävää ja kokeessa 4 muunneltua malliinvertaustehtävää. Erottelukynnykset estimoitiin sovittamalla psykometriset funktiot dataan. Keskiarvoistamisessa hyödynnettyjen näytteiden määrä estimoitiin ekvivalentin kohinan mallinnuksella.

Erottelukynnykset nousivat ulkoisen kohinan noustessa, mutta laskivat elementtien määrän kasvaessa, parannuksen ollen suurempaa korkeammalla kohinalla. Havainnoijan käyttämien näytteiden mallintaminen saatavilla olevien näytteiden potenssina sopi erinomaisesti dataan. 64:än elementin ärsykkeelle estimoitu efektiivisten näytteiden määrä vaihteli välillä 16 ja 41. Kontrollikokeet vahvistivat, että suoritus parani elementtien lukumäärän, ei ärsykkeen pinta-alan mukaan. Jopa voimakkaasti vinolla sävyjakaumalla suurin osa havainnoijista päätyivät yksinkertaiseen keskiarvoistusstrategiaan estimoinnissa. Vähemmän suoraviivaisella vertailutehtävällä tulokset näyttäytyivät vähemmän johdonmukaisina.

Havainnoijat ottavat näytteitä ja keskiarvoistavat ärsyke-elementtejä arvioidakseen keskimääräisen värisävyn, samoin kuin muiden alueiden joukkohavainnoinnissa. Havaittu näytteenotto ylittää selvästi aiemmat estimaatit viitaten globaaliin otantaan, ja suoritukseen vaikuttaa eniten pinta-, ei reuna-informaatio. Yksinkertainen keskiarvoistus säilyi havainnoijien spontaanina strategiana ei-normaalilla sävyjakaumalla. Tulokset olivat myös riippuvaisia tehtävän yksityiskohdista, tähdentäen varovaisuutta verrattaessa erityyppisiä kokeita.

Avainsanat – Nyckelord – Keywords
Värisävyn keskiarvoistus, joukkohavainnot, värinäkö

Säilytyspaikka – Förvaringställe – Where deposited
Helsingin yliopiston kirjasto – Helda / E-thesis (opinnäytteet) *https://ethesis.helsinki.fi/*

Muita tietoja – Övriga uppgifter – Additional information

Table of contents

1. Introduction

Our perceptual world is rich with different color sensations. Besides the main color categories, we are able to discriminate small differences in a multidimensional color space. The underlying mechanisms of color perception have attracted extensive research (see e.g. Gegenfurtner & Kiper, 2003, for a review). Although some studies have used more complex stimuli (e.g. Kimura, 2018; Maule & Franklin, 2015; Milojevic, Ennis, Toscani & Gegenfurtner, 2018; Olkkonen, Hansen & Gegenfurtner, 2008), much of the knowledge in human color perception so far relates only to uniformly colored fields, such as the highlighted color squares in Figure 1.

In natural scenes color hardly ever exists in such discrete uniform fields. Instead, the color appearance of a single object is formed from a chromatic distribution. Partly this is due to the object's surface reflectance or how different wavelengths of light are reflected from the object's surface, and partly due to non-uniform illumination. Consider the apples in Figure 1 and the range of different colors highlighted in the picture. Despite this variety of colors in the retinal image we have no difficulty determining that these apples are red and that the left one is the redder of the two. What is this decision based on? Is there some selection and pooling of information? Do we take in some or all of the available information? Do we simply average out the surface color or are there even more complicated processes at work?

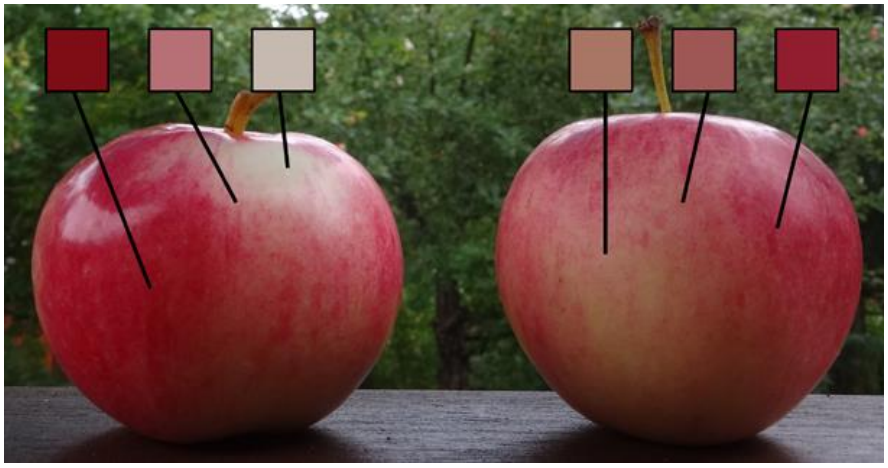

Figure 1. *A photograph of two apples with colors in different locations highlighted. Despite the large range of different surface colors, most observers have no trouble assigning a representative color for these apples.*

As of yet, relatively little is known about how perceived color is determined in these more complex, natural viewing conditions (but see: Witzel & Gegenfurtner, 2018). The subject of this study is how the perception of color is formed over spatial fields of varying color information. The focus is solely on the color of the surface texture of the objects, omitting the additional complexities that different lighting conditions bring to the picture, such as shadows and highlights (Figure 2a, left). For the sake of illustration, if we assume completely uniform illumination, the color variation of the object is reduced to the variation in surface reflectance (Figure 2a, right). This surface reflectance is a property of the object (instead of a perceptual property) and what we will refer to as the object's surface color in this study. For clarity, keep in mind that the signal reaching the observer's retina still originates from the source of illumination, reflected from the object's surface. This will be elaborated on in the next section. However, the question remains, does the visual system form an average of the object's color as in Figure 2b and how?

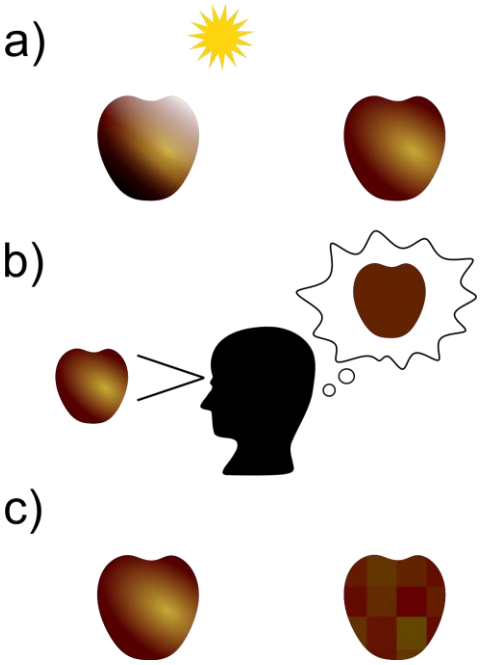

Figure 2. *Issues in estimating surface color. a) In natural settings, the directional source(s) of illumination commonly cause highlights and shadows as in the left apple. With the theoretical apple on the right, illumination is uniform from all directions. b) As we can estimate an average color of many everyday objects, does our visual system form a uniform color representation for the object? c) On a highly variable surface such as an apple, it is difficult to measure or control the variability of the color signal. For experimental purposes, the color variation can be broken down into clear and easily controllable elements with differing color signals.*

Of main interest in the context of this work are the effectiveness and extent of available information used in forming an overall color percept. To accurately control the variation in color, the stimuli are divided into distinct elements as illustrated in Figure 2c. Additionally, we extend upon earlier studies by systematically testing how spatial manipulations affect the overall color percept. Namely, whether having the color elements form a unified stimulus differs from spatially separated color elements, or if the signal surface area or overall stimulus area exert an effect. Also, we examine

whether the observers always use a simple averaging strategy by testing discrimination with highly skewed color distributions. Finally, we explore discrimination in a different comparison task that allows including the complete range of hues.

1.1. Color vision, color spaces and hue

To fully understand the premise of the current research, one must have basic understanding of human color vision. The current section gives the background information for understanding the reasoning behind the research methodology. The current chapter is not directly motivating the research question of this study and readers familiar with the basics of color vision may wish to skip directly to the next section.

Human color vision is based on three cone photoreceptor types in the retina that are differentially sensitive to different wavelengths of light (Stockman & Brainard, 2010). Long before anything was known of retinal cell types, researchers were able to determine that color vision is based on a combination of only three signals, commonly known as *trichromacy* (The Young-Helmholtz thrichromatic theory, see e.g. Balaraman, 1962). Also based on purely empirical observations, the second stage of processing color information was found to depend on two chromatic antagonistic channels (commonly, but inaccurately referred to as red-green and blue-yellow) and one achromatic (white-black) channel (Hering's opponent process theory, popularized by Hurvich and Jameson, see e.g. Hurvich & Jameson, 1957). In short, the neural coding of color is economical as it represents the whole spectral variation as the balance of just three signals. This manner of neural encoding also clearly determines the perceptual limits of human color vision.

The understanding of trichromacy enabled the development of *colorimetry* (e.g. Brainard & Stockman, 2010), which serves to link physical measures of light to perceptual attributes of color. The first stage of human color vision simplifies the complex light input into three neural signals. To elicit two identical color percepts, all that is needed is to match the three cone activations. Thus, many widely different wavelength profiles can generate the same color perception as long as the relative absorption rate for the three cone types remains similar. Such different lights that are indistinguishable to a human observer are called *metamers* and they were an essential tool to delineate the quantitative relations between light and perception.

The fact that color is represented in the human visual system by only three signals also means that only three primaries (e.g. monochromatic lights) are necessary to replicate the whole range of

human color percepts (Smith & Pokorny, 2003; Wyszecki & Stiles, 1982). In color matching experiments, observers create metamers by matching a proportional mixture of three primaries to another light with a known wavelength (Figure 3a). When this process is repeated throughout the visible spectrum in finely spaced steps, one can create *color matching functions* for the used primaries (Figure 3b). Color matching functions signify the power of each primary needed to replicate any monochromatic point in the visible spectrum. The color matching functions can also be translated to another set of primaries by a mathematical transformation. The aforementioned principles are effectively utilized in reproducing color in print and on monitors.

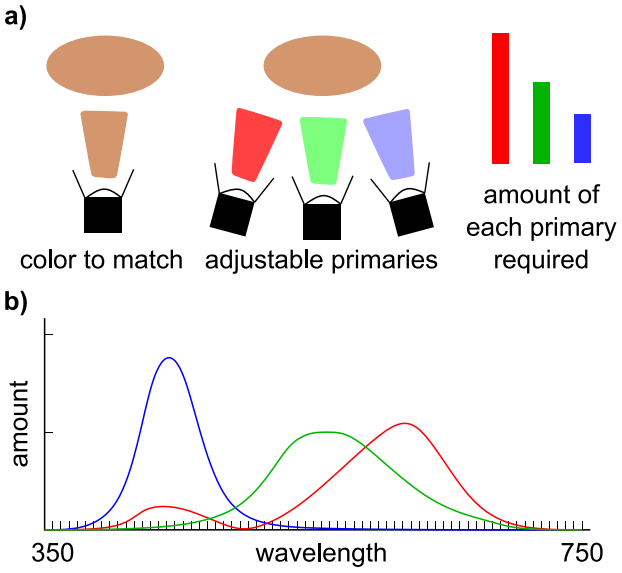

Figure 3. *Color matching experiments. a) A common procedure in color matching experiments. The observer can control the intensity of three lights (primaries) and adjusts them to match to a light of particular wavelength. The intensities are recorded for lights throughout the visible spectrum in finely spaced steps. b) Based on the recorded intensities, color matching functions can be calculated. Please note that the one in this figure is hypothetical and for illustration purposes only.*

To establish a standard set of color mathing functions for the average observer, Commission Internationale de l'Eclairage (CIE) based its estimates on the results of the color matching experiments of Wright (1929) and Guild (1931). The resulting CIE RGB system was further developed into the more convenient CIE XYZ system, also know as the CIE 1931 standard colorimetric observer (Broadbent, 2004; Fairman, Brill & Hemmendinger, 1997), which is still a widely used reference standard.

CIE XYZ represents a *color space*, an organization or mapping of color (Fairchild, 2013). There are many different color spaces to fit different purposes in color reproduction in different media etc. There are also other color spaces developed to encompass the range of human color perception, but aiming to portray different aspects more conveniently. For example, CIELUV and CIELAB aim for perceptual uniformity, meaning equal discriminability with equal distances anywhere in the color

space (Robertson, 1977). DKL (Derrington, Krauskopf and Lennie) color space uses the estimated opponent channels as its axes to emphasize the opponent process in color vision (Derrington, Krauskopf & Lennie, 1984).

In addition to providing a clear mapping of different colors, standardized color spaces have the benefit of being interchangeable (with certain limitations, such as range). Basically, as the color space is defined with clearly specified metrics, the coordinates can be manipulated mathematically (Brainard & Stockman, 2010). Mathematical transformations can be used to move from one color space to another. For example, in this study it enables defining a stimulus in relevant perceptual metrics and transforming those into physical metrics needed for accurate display of the stimulus.

The common perceptual metrics for color are *hue*, *chroma* and *lightness* (Fairchild, 2013). Hue has a commonsense interpretation as the balance of red, green, blue and yellow the color possesses. Hue forms a perceptually circular metric, so it has no natural zero point, nor are hues numerically comparable besides in relation to some arbitrary point. Chroma is the colorfulness (or richness of color) as a degree of perceptual difference from grey of the same lightness. For the sake of future reference, a closely related, but slightly different attribute to chroma, is *saturation* (colorfulness in relation to its brightness). Lightness is the perceived brightness compared to maximal brightness in given lighting. It is important to note that these are relative metrics, which are more informative than absolute metrics in changing environments.

When we move from the color percepts of light into estimating object surface color, it is essential to keep in mind that the light hitting the eye results from a combination of illumination and reflectance. The source of the signal in the environment is the illumination with surface reflectance only manipulating this signal. Thus, changes in illumination change the signal that is reflected from the object to the retina drastically. This creates a constancy problem. As illustrated in Figure 4, the same white shirt gives us a very different retinal image under two different lights. Still, we somehow consider the shirts to possess the same surface color (color constancy, see e.g. Foster, 2011).

When we add the whole complexity of a natural viewing environment, many (seemingly) more complex color vision mechanisms emerge (Witzel & Gegenfurtner, 2018). Effects produced by color constancy (discounting the effects of illumination when interpreting colors, Brainard, Cottaris & Radonjić, 2018; Foster, 2011), color memory (Hansen, Olkkonen, Walter & Gegenfurtner, 2006; Olkkonen, Hansen & Gegenfurtner, 2008) and color categories (Witzel & Gegenfurtner, 2015; 2016) assist in creating a more stable and predictable visual presentation of our environment

(Witzel & Gegenfurtner, 2018). Color is also not processed separately from form as previously thought (Gegenfurtner & Kiper, 2003; Shapley & Hawken, 2011; Rentzeperis, Nikolaev, Kiper & van Leeuwen, 2014), affording interplay between different scene and object properties.

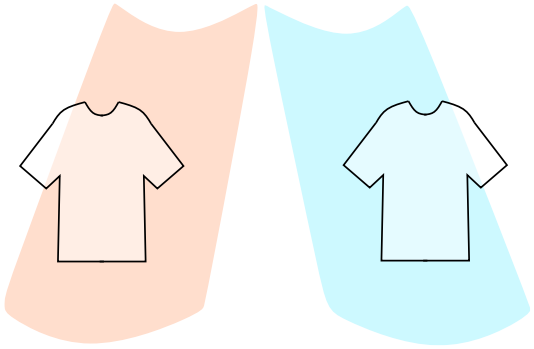

Figure 4. *An illustration of color constancy. In natural settings we face different kinds of illumination. Different illuminations cause object surfaces to reflect a very different signal to the retina, as with these theoretical white shirts. Still, our perception of objects' surface colors is not wholly inconsistent.*

All in all, color percepts are formed in a complex interaction of environmental properties and neural mechanisms. Even with the many complex mechanisms of color vision, the way percepts of object surface color are formed when there is variable information, is a meaningful question by itself. Even if there was no constancy problem, the question of spatial integration would still stand. To gain meaningful insights into particular processes in color vision requires rigorous control of many external factors. Therefore, the more complex effects of perception in natural viewing are purposefully left out of the equation at this stage of the research.

1.2. Combining information

If the visual system is to properly estimate the average color of the apples in Figure 1, it has to combine color information from different spatial locations of the apples. More often than not, we have more than one source of information available to us, such as binocular disparity and texture when estimating surface slant. Integrating information from multiple cues shows improved performance for a multitude of tasks both intramodally (e.g. Hillis, Watt, Landy & Banks, 2004; Knill & Saunders, 2003; Landy & Kojima, 2001; Young, Landy & Maloney, 1993) and intermodally (e.g. Alais & Burr, 2004; Ernst & Banks, 2002; Hillis, Ernst, Banks & Landy, 2002). This is called cue integration (Landy, Maloney, Johnston & Young, 1995). Furthermore, a particular form of cue integration swiftly integrates information from several parts of a stimulus or perceptual field as overall statistical information called *ensemble percepts* (for reviews, see: Bauer, 2015; Whitney & Yamanashi Leib, 2018).

Ensemble perception refers to the way our perceptual system stores information from large groups in particular situations, for example when seen at a glance. Ensemble percepts are characterized by observers having accurate knowledge of the overall statistical regularities, but poor ability to accurately recall individual set members (Ariely, 2001). Ensemble perception applies from averaging low-level features such as direction of motion (Watamaniuk & McKee 1998; Watamaniuk, Sekuler & Williams, 1989) and orientation (Dakin, 2001; Parkes, Lund, Angelucci, Solomon & Morgan, 2001) to averaging high-level features such as facial emotion (Haberman & Whitney, 2007; 2009) and facial identity (de Fockert & Wolfenstein, 2009). The representations are not always independent of other visual aspects and may correlate with features that are on a similar level of visual processing, such as color and orientation or facial identity and facial expression (Haberman, Brady & Alvarez, 2015). There also seems to be some domain specificity in different types of statistics, as simultaneously estimating the mean and numerosity from the same set does not impair precision, whereas estimating the same statistic from two different sets does (Utochkin & Vostrikov, 2017; see also: Yang, Tokita & Ischiguchi, 2018).

To summarize the basic principles of ensemble perception Whitney and Yamanashi Leib (2018, p. 112) offer an operational definition:

- *Ensemble perception is the ability to discriminate or reproduce a statistical moment.*
- *Ensemble perception requires the integration of multiple items.*
- *Ensemble information at each level of representation can be precise relative to the processing of single objects at that level.*
- *Single-item recognition is not a prerequisite for ensemble coding.*
- *Ensemble representations can be extracted with a temporal resolution at or beyond the temporal resolution of individual object recognition.*

Although it is generally accepted that ensemble perception enhances visual cognition (Alvarez, 2011), it is still relatively unknown how the visual system manages this swift and complex operation. Theories have ranged from limited sampling in selective attention (Myczek & Simons, 2008) to a more global process of distributed attention (Ariely, 2001; Chong & Treisman, 2003). Many studies have estimated the number of samples observers can utilize assuming a sampling strategy. When processing each item in a set individually, not all of the items in the stimulus can be taken into account because of limited attentional (Scimeca & Franconeri, 2015) and working memory resources (Luck & Vogel, 2013). In limited sampling, the visual system would choose a limited amount of available items and base the estimation on them. Combining several results,

Whitney and Yamanashi Leib (2018) suggested that the average sampling rate would be approximately the square root of the items available.

If each sampled item was processed individually, the square root of items in larger sets would clearly exceed the limits of attention and working memory. For example, the maximum capacity of working memory was long considered to be approximately four items (Cowan, 2010; but see: Ma, Husain & Bays, 2014; Cowan, 2017). More recent investigations into sampling in ensemble perception have not only pointed towards a more global sampling in distributed attention with larger sets, but also differences in strategy with smaller sets (Tokita, Ueda & Ishiguchi, 2016). Lau and Brady (2018) propose that neither holistic parallel processing nor limited sampling strategies by themselves manage to explain observers' performance in different variability estimation tasks. Instead, they suggest that there are multiple, sometimes indirect strategies and estimation heuristics, which also depend on various task details. One such strategy involves smart subsampling (directing sampling to the most meaningful targets) and a simple heuristic. Another strategy makes use of consistent low-level information in the stimulus.

The use of low-level information supports the idea of texture-like processing of the visual stimulus, suggested by Im & Halberda (2013). Also in support of texture-like processing of visual ensembles is that the same neural structures are found to be primarily involved with processing both textures and ensembles (parahippocampal place area, PPA: Cant & Xu, 2012; Cant & Xu, 2014; Cant & Xu 2017). This area is also central to processing of scenes (Epstein & Kanwisher, 1998) and seeing the gist of a scene has been tied to ensemble perception (Oliva & Torralba, 2006) and ensemble texture representations (Brady, Shafer-Skelton & Alvarez, 2017). There seems to be at least some connection between texture processing and ensemble perception (Victor, Conte & Chubb, 2017), but it is unclear what this connection implies for the dimension of color. There is no clear consensus on the exact relation between texture and color. The two aspects seem both closely related (Saarela & Landy, 2012; 2015) but also separable (Cant, Large, McCall & Goodale, 2008; Cavina-Pratesi, Kentridge, Heywood & Milner, 2010). Therefore it is uncertain whether the achromatic and chromatic mechanisms are truly similar in nature.

1.3. Previous research in color ensembles

Just as in other domains, the average hue can be estimated as a statistical moment from a collection of different hues (Maule, Witzel & Franklin, 2014). Observers are able to accurately determine a mean color percept for a collection of briefly presented hues and, as is common with ensemble

perception, the mean hue does not necessarily need to be represented at all in the stimulus. The average hue also seems behaviorally meaningful. The colorimetric mean hue is a very good predictor of color categorization in a naturalistic task of sorting autumn leaves by color, as shown by Milojevic, Ennis, Toscani and Gegenfurtner (2018).

However, there are additional complexities in estimating color averages. In a task to match a uniform color to a 20 x 20 color mosaic (randomized from 9 preset colors) by method of adjustment, Kuriki (2004) found that observers drifted away from the colorimetric mean as the variation of colors increased. Instead, the percept was strongly biased towards the most saturated color in the mosaic when all the elements were roughly within the same color category. In a similar matching by adjustment task, Kimura (2018) found that the color variation played a crucial role in color averaging. However, the bias towards the most saturated color persisted even with smaller color variation, while the mean hue of the matches closely followed the mean hue of the mosaics. The author reasoned that there is differential processing of hue and saturation information, also suggesting that the most realiable information about surface reflectance comes from the most saturated colors in natural images, thus giving rise to this bias. The effects of saturation have also been studied in relation to unique hues which are defined as hues that observers' perceive as pure (such as pure red), containing no traces of other hues (Mollon & Jordan, 1997). Sunaga and Yamashita (2007) found that observers' estimates of the mean color of a random-dot mosaic formed from two colors only differing in saturation followed more closely the locations of unique hues rather than the colorimetric mean. This indicates that the perception of the whole is formed by integration on the level of individual color percepts, not at the level of color metrics.

The averaging of hues can still be studied from the ensemble perception viewpoint when the confounding effects of saturation are controlled. An obvious further question stands as in other ensemble perception studies, how is hue information pooled or sampled from the stimulus? By simulating a theoretical observer and comparing to data, Maule and Franklin (2016) came to the conclusion that sampling just two elements of the ensemble was sufficient to account for the performance of real observers in estimating the mean hue of an ensemble. However, there are some limitations to their findings. The hues in their stimuli were selected from a predetermined set of clearly discriminable hues, which made the stimuli very different from the more gradual hue distributions in everyday objects. Their stimuli also contained a maximum of four different hues. Although the stimuli included a total of 16 elements, their previous results showed that the amount of color cues does not affect performance if they only contain more of the same hues already included in the ensemble (Maule & Franklin, 2015). Therefore, the task didn't significantly differ

from a situation with just four elements of information available to the observer. Finally, their simulations only included either early noise (at the stage of processing individual elements) or late noise (at the stage of integration and forming a decision) (Maule & Franklin, 2016). However, an equally if not more likely scenario is that there is signal noise at both of these stages.

Besides the mean, human observers can also extract summary information of the variance and of the shape of a distribution of hues. The experiments were based on a finding of Corbett and Melcher (2014) that stable statistical representations facilitate visual search. Chetverikov, Campana and Kristjánsson (2017) employed a visual search task (pop-out priming) in a hue ensemble. Based on observers' reaction times, they found that observers could utilize knowledge of prior hue distributions, including variance and skewness. Observers were also found to learn new distributions quickly and that robust learning only occured on larger ensemble sets. Yet there is a danger of overgeneralization with the results of Chetverikov, Campana and Kristjánsson. Their experiment did not directly measure the formation of ensemble percepts. It is therefore unclear what type of information the observers utilized. Just perceiving and exploiting statistical properties such as outliers does not necessarily indicate that the distribution was coded into a statistical summary.

All in all, it is known that some summary statistics of hue ensembles can be formed swiftly from even briefly presented stimuli and this is in line with results from other visual domains such as size and orientation. The simple averaging of hue does not generalize to the saturation of colors. However, the precise integration process in color and its dependent factors are still rather poorly understood. There has been a very limited amount of studies on this specific topic and they share some general limitations. Firstly, in the averaging experiments, hues have always been drawn from a limited, predetermined set of hues. Such conditions, especially when the different hues are clearly distinguishable, are in sharp contrast to the more continuous surface colors in naturalistic settings. Second, there has been limited focus on the effects that the spatial properties of the stimuli have, such as distance between elements, element size and whole stimulus size. Third, the tasks in the previous experiments have been very similar, most commonly matching a uniform stimulus to a stimulus with a chromatic distribution. As the particular strategy of the observers depends on the task (Lau & Brady, 2018) it is unclear how the previous results generalize to different tasks and experimental conditions.

1.4 The current study

In this study we aim to address the limitations in previous research for a more comprehensive understanding of how hue ensembles are processed by the visual system. This study applies psychophysical methods and modeling to uncover details of human information processing while making no implications to the physical location or structure of said processes. In contrast to the predetermined sets of stimulus hues in previous studies, we employ a continuous hue distribution. The study consists of four separate experiments, each focused on a different aspect of hue averaging.

Experiment 1 focuses on observer performance as a measure of how well observers can utilize increasing amounts of information (stimulus elements). We manipulate ensemble set size and external noise in a simple two-interval forced choice (2IFC) comparison task ("yellower" or "bluer") in a limited hue range (between yellow and blue). Equal variance modeling (further elaborated in methods) is employed to gain insights into how observers extract summary statistics. We estimate how many individual samples (individual elements) observers need to average to account for their level of performance. In addition, a separate manipulation explores if there is a difference in hue averaging when the elements are not connected but spatially separated from each other. Theoretically, connected elements could be seen as a single object whereas the disconnected elements are more likely to be seen as separate objects. The difference in object individuation could therefore affect hue averaging.

In Experiment 2, spatial manipulations of the stimuli are extended from Experiment 1. Our goal is to explore if the number of hue elements is the only factor contributing to hue averaging in our experiments, as Experiment 1 confounds stimulus size and number of elements. The manipulations account for the number of elements, the signal area of the elements and the total surface area that the elements extend to, additionally including a wider separation of elements than in Experiment 1. Varying one aspect while controlling the other two, we can determine which one(s) determine performance. The task in Experiment 2 is otherwise similar to the one employed in Experiment 1.

For Experiment 3, the question is whether changes in the type of noise would produce changes to observers' *spontaneous* averaging strategy. A small set of set size conditions are selected and a strongly skewed distribution of hues is used. Otherwise the task is again similar to Experiment 1. Instead of performance, the appropriate measure is the possible bias in responses. Using a skewed distribution generates two extreme propositions for averaging strategy. If the observers weigh all elements equally, their responses should be centered at the distribution mean. However, if observers

employ a different strategy, such as discounting outliers or emphasizing most numerous hues (centered at the mode), their responses should be biased towards the mode of the distribution.

Experiment 4 attempts to extend and generalize Experiment 1 to the whole range of hues available to us. While the task in Experiment 1 necessarily limits the range of hues between yellow and blue, Experiment 4 employs a task of comparing the test stimulus to two uniform comparison stimuli (modified match-to-sample task). This task allows meaningful comparisons in the complete range of hues. However, the actual comparison is slightly different from Experiment 1 in which two hue averages are compared. Experiment 4 also includes a full factorial design to explore whether the results with disconnected elements would generalize to a wider range of different ensemble set sizes.

2. General methods

The four experiments conducted in this study were very similar in their methodology. The general methods apply to all aspects of the experiments unless otherwise mentioned.

2.1. Observers

12 observers (including the author and his supervisor) took part in the experiments (8 female, age range 20-60, mean age 33 years). Their participation in different experiments is listed in Table 1. The observers were given a single identification number to carry over all of the experiments. The interested reader can therefore compare observers' performance across different experiments. Besides the author and his supervisor, all other observers were naive to the theoretical aims of the experiment. All observers had normal or corrected-to-normal visual acuity and normal color vision as assessed by Ishihara plates (Ishihara, 1973). Observers gave informed consent and received movie ticket vouchers for their time.

Table 1. *List of observer participation in different experiments.*

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10	O11	O12	total
Experiment 1	x	x	x				x	x	x	x	x	x	9
Experiment 2	x			x	x		x	x	x	x	x		8
Experiment 3	x	x	x	x		X	x			x	x		8
Experiment 4	x			x			x	x		x		x	6

2.2. Apparatus

The experiments were conducted on a HP Z230 Desktop PC running Matlab with PsychToolBox extensions (Brainard, 1997; Kleiner et al., 2007). Stimuli were presented on a 23 inch ViewPixx monitor controlled by a Nvidia Quadro K620 graphics card. The monitor resolution was 1920 x 1080 pixels, with 100 Hz refresh rate, 10-bit color channels and a maximum luminance of 250 cd/m². The monitor was calibrated using X-Rite i1Pro spectrophotometer with standard methods (Brainard, 1989) and the white point to match D65 standard illumination. Observers took the experiment in a dimmed room and their viewing distance was held constant at 90 cm from the screen using a chin rest. Observers gave their responses using a regular keyboard.

2.3. Stimuli

For all color stimuli in the experiment, hues were picked from a hue circle in CIELAB color space with monitor white point as the reference (Figure 5). Lightness was held constant at a value of 60 and saturation at 50. Thus, all presented colors except the background only differed by their hue. In order to have the color stimuli pop out slightly better, the background was a uniform gray with lightness set to 50. The different hues were represented numerically by their corresponding angle on the hue circle. The mean color of the test stimulus was randomized for each trial between 140° and 160° of angle on the hue circle. The hue range of the stimuli was therefore limited between the yellow and blue category limits, to keep the stimuli relevant to the task of discriminating whether the comparison stimulus was “yellower” or “bluer” than the test stimulus.

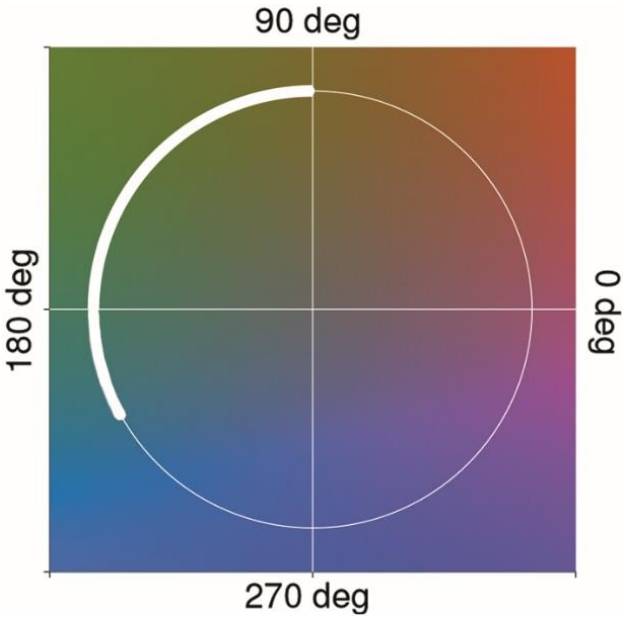

Figure 5. *An equiluminant plane in CIELAB color space. A section of the color space at a particular L (lightness) level. On x-axis is the a-factor (approximating red-green) and on y-axis the b-factor (approximating yellow-blue). The distance from the origin determines the chroma. In Experiments 1, 2 and 3, the total hue variation was approximately in the highlighted area. In Experiment 4, the whole hue circle was used. The colors represented here do not exactly match the ones produced in our experiments and are for illustration purposes only.*

The test stimulus consisted of a square grid with a varying number of square elements (from 1 to 64) near the middle of the screen. The stimulus midpoint was randomized around the screen center within a 60 pixel range to a random direction, in order to keep observers from fixating the same part of the stimulus over numerous repetitions and to lessen the effects of visual adaptation. Each element was filled with a uniform color and extended 1 degree of visual angle. Depending on the experimental condition, the elements were either connected or separated. For no-noise trials, all stimulus elements shared the test stimulus mean hue, making the stimulus uniform in hue. For noise trials, a set of random hue angles for the elements were drawn from a circular von Mises distribution centered at the test stimulus mean hue. The von Mises distribution's probability density function for angle *x* is of the form:

$$d(\mathrm{x}|\mu,\kappa) = \frac{e^{\kappa\cos(\mathrm{x}-\mu)}}{2\pi I_0(\kappa)} \qquad (1)$$

Where $I_0(\kappa)$ is the modified Bessel function of order 0 and μ and $1/\kappa$ are analogous to μ and σ^2 of the normal distribution. For low noise conditions $\kappa = 40$ and for high noise conditions $\kappa = 15$ values were used.

The comparison stimulus was similar to the test stimulus besides having its mean hue from one of the 9 comparison levels around the test stimulus mean (see Figure 7) and an independent noise sample. The comparison levels were selected from preset ranges for each noise (e.g. Experiment 1) or set size (e.g. Experiment 2) condition. The decision was based on the practice results of

individual observers so that the comparison stimulus covered the range that allowed measuring thresholds – the task was not too easy or too difficult.

2.4. Procedure

The observer was instructed to assess the perception gained by the whole stimulus of the test and the comparison and to evaluate their "representative" hue. The word "mean" was not used in the instructions in order to avoid encouraging particular strategies. The observer's task was to respond whether the hue represented by the whole of the comparison stimulus was "yellower" or "bluer" than the one for the test stimulus. The observer responded by pressing the left or right arrow key on the keyboard. The directions indicating responses were switched between each observer.

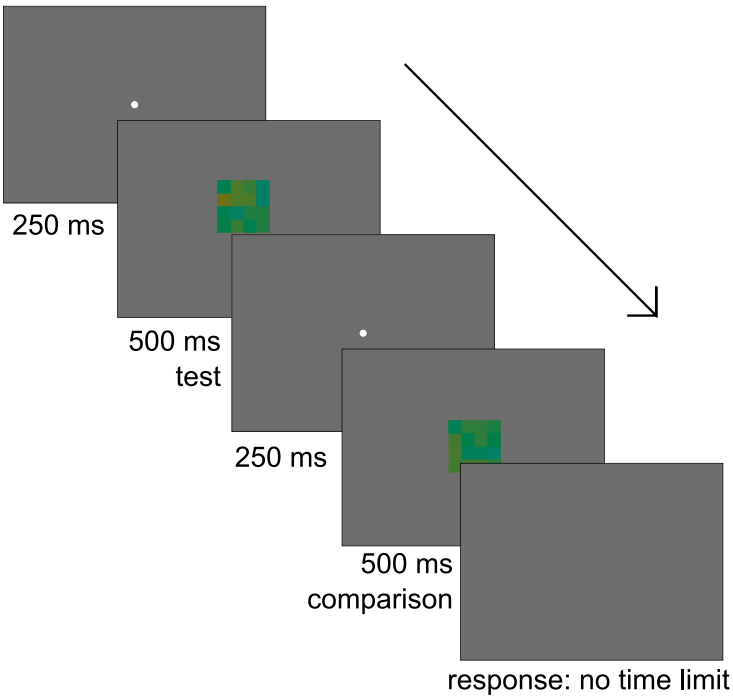

Figure 6. *The course of a single trial. The observer is shown a dot indicating the center of the screen for 250 ms, then the test stimulus for 500 ms. This is followed by an inter-stimulus interval with the center dot for 250 ms and then the comparison stimulus for 500 ms. Finally, the screen is blank until the observer gives a response and the next trial begins.*

In each trial (Figure 6) the observer was shown a blank screen (uniform gray of the background) with a white dot of 0.1 degrees of visual angle indicating the middle of the screen for 250 ms. The observers were not instructed to fixate the white dot, only told that the dot would serve as an indicator around which the stimulus would appear. Next, the test stimulus was displayed for 500 ms. Again, the blank screen with a center dot was displayed for 250 ms, now followed by the comparison stimulus for 500 ms. Finally, the screen was blank until the observer gave a response. The observer received no feedback. Test and comparison intervals were not randomized (always

presented in a fixed order) for consistency across experiments, as randomization was not feasible in some experiments.

2.5. Analysis

Analysis of the data was performed in Matlab (version R2018b, build 9.5.0.944444), except for the analyses of variance (ANOVA) which were done in R (version 3.5.1). Each comparison level in an experimental block served as a data point for the number of "bluer" responses by the observer. A cumulative Gaussian psychometric function (PMF) was fit to this response data with mean and standard deviation (SD) as parameters. Mean described response bias from the true value (used in Experiment 3) and SD described discrimination threshold (used in Experiments 1, 2 and 4).

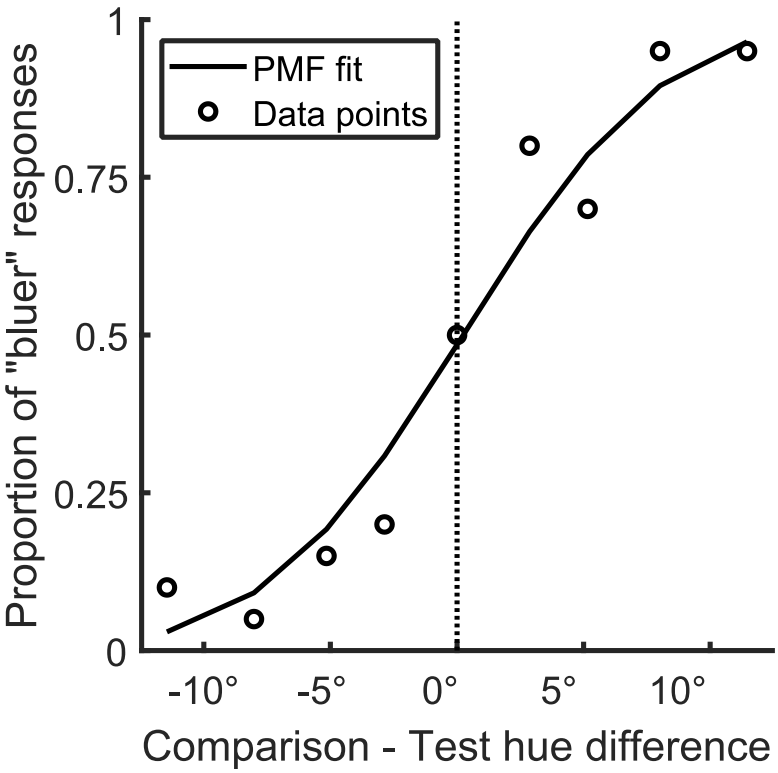

Figure 7. *Fitting a psychometric function. On the x-axis is the difference in hue angle between the test and comparison stimuli means, with the test stimulus mean at the zero point. On the y-axis is the proportion of "bluer" responses by the observer and these are indicated by the circles for each of the 9 comparison levels. A cumulative Gaussian psychometric function is fit to the data points and the obtained fit parameters serve as our dependent variables.*

A bootstrap method was used to estimate the standard error of mean (SEM) for individual observers in different experimental conditions (Wichmann & Hill, 2001). The data were sampled with replacement for each data point and a PMF was fit to the sampled data. This was repeated for 2000 iterations and the 68.27% confidence limits of the resampled parameters were drawn to represent

+/- 1 SEM. Repeated measures ANOVAs were also performed over all observers to test the effects of the particular experimental manipulations in each experiment.

2.6. Modeling

For experiment 1, *equivalent noise modeling* was employed to estimate the limits of observers' effective sampling (cf. Dakin, 2001). The equivalent noise model can be considered a special case of cue integration when the strategy to estimate an ensemble or group mean is simple averaging with equal weights, following the assumption that internal noise affects all cues equally. The model also assumes limited capacity in sampling – not all samples can necessarily be used. By introducing external noise and estimated internal noise to the equation, one can estimate how many individual cues are needed to average out the noise to match the observer's performance. The basic form of the model used was:

$$\sigma_r = \sqrt{\frac{\sigma_i^2 + \sigma_e^2}{n_s} + \sigma_o^2} \qquad (2)$$

Where:

- σ_r is the noise in a single stimulus, i.e. the total uncertainty in observer's estimate of one of the two presented stimuli. Multiplying by $\sqrt{2}$ equals the performance measured from observers or PMF SD.
- σ_i^2 is internal noise variance, which is assumed equal for each hue element.
- σ_e^2 is external noise variance, in this case controlled by the experimenter and applied as variability in element hues.
- n_s is samples effectively utilized by the observer which is of our main interest in model estimation. Larger values of n_s result in smaller uncertainty in the observer's responses and thus better performance.
- σ_o^2 is various other sources of noise such as noise in the integration process.

In our experiment, values of σ_r were set by observer performance and used as the target of the fit. σ_e^2 was also set by the external noise in the particular experimental condition. The free parameters therefore included σ_i^2, σ_o^2 and n_s. n_s was additionally limited to a maximum equal to the number of elements in the stimulus (noted as n_e), as a larger amount would have been theoretically unsound.

The model was fit to the data and best-fitting parameters were estimated by maximizing log-likelihood for two different variations. This was done for each observer individually. In the first variation of the model we estimated a single set of values for the free parameters for each observer. In the second variation of the model, instead of estimating a value directly for n_s, we estimated a value for k ($0 < k < 1$), so that $n_s = n_e^k$ and always a *proportion* of stimulus elements. These two models will be referred to as the simple model and the power model, respectively. Besides testing the model fit individually for each observer, the model fit was also tested for observers' averaged results. However, with averaged results, the model was fit to the averaged discrimination thresholds instead of the raw data.

3. Experiment 1

3.1. Specific methods

Experiment 1 focused on observer performance with increasing amount of information or stimulus elements. In the main experiment of Experiment 1, conditions formed a quasi-factorial design (illustrated in Figure 8) in which the number of elements (1, 4, 16, 64), distance between elements (connected, disconnected by a distance of 1/3 element size), and amount of external noise in the hue of elements (no noise, low noise, high noise) served as independent variables (see Figure 8). The distance between elements was varied only for the 16-element stimulus. Also, the 1-element stimulus was only tested with the no-noise condition to avoid redundancy. This resulted in 13 experimental blocks, which were repeated twice in a random order for a total of 26 experimental blocks.

Experiment 1 consisted of two parts: practice and the main experiment. A short demo introducing the different stimuli and the task was shown before practice. Practice runs contained a fixed subset of 9 relevant experiment conditions with a small number of repetitions. A practice run was conducted at the beginning of each measurement session. In Experiment 1 practice runs, there were 5 repetitions for 9 comparison levels for the 9 experimental conditions included, resulting in 405 practice trials total. For the main experiment of Experiment 1, a single trial block contained 10 repetitions for the 9 comparison levels. This means that observers completed 90 trials in each of 26

blocks, 2340 trials in total for the main experiment. The whole Experiment 1 took approximately 2 hours to finish, which observers completed in one or two sessions.

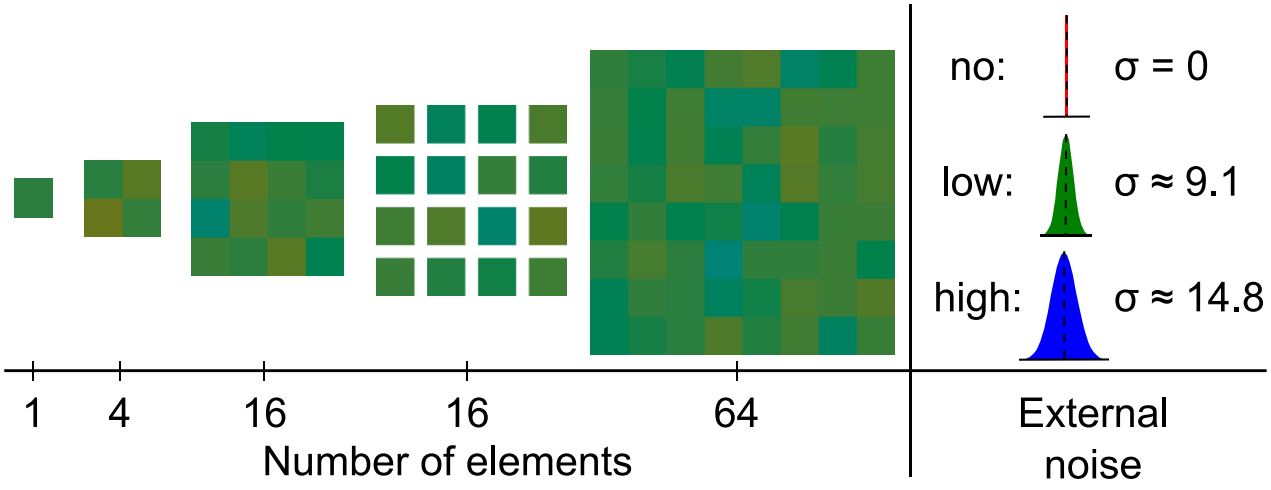

Figure 8. *Experimental conditions for Experiment 1. The experiment included five different types of stimuli as seen on the left side of the figure and three levels of added noise for the element hues on the right. The 1-element stimulus was only tested in the no-noise condition to avoid redundancy. Please note that these stimulus examples in print do not exactly match with the stimuli produced on monitor in the experiments. The experimental stimuli were also not displayed on a white background. These example stimuli are for illustration purposes only.*

3.2. Results

Of the nine observers in Experiment 1, one reported not being able to tell shifts into "bluer" and "yellower" in the task. This difficulty persisted even after being shown a continuum from yellow to blue. For some of the experimental conditions a PMF could not be fit, so this observer was omitted from further analysis. The eight remaining observers' results for all experimental conditions excluding the disconnected condition can be seen in Figure 9. There are three consistent trends visible: 1) the task is more difficult with more external noise 2) performance improves as the number of elements increases 3) the improvement in performance with increasing number of elements is more pronounced with higher noise levels. These effects are also seen in the two-way repeated measures ANOVA with the main effect of external noise ($F(2, 14) = 69.24$, $p < .001$), set size ($F(2, 14) = 98.95$, $p < .001$) and the interaction of the two ($F(4, 28) = 20.11$, $p < .001$) all being statistically significant.

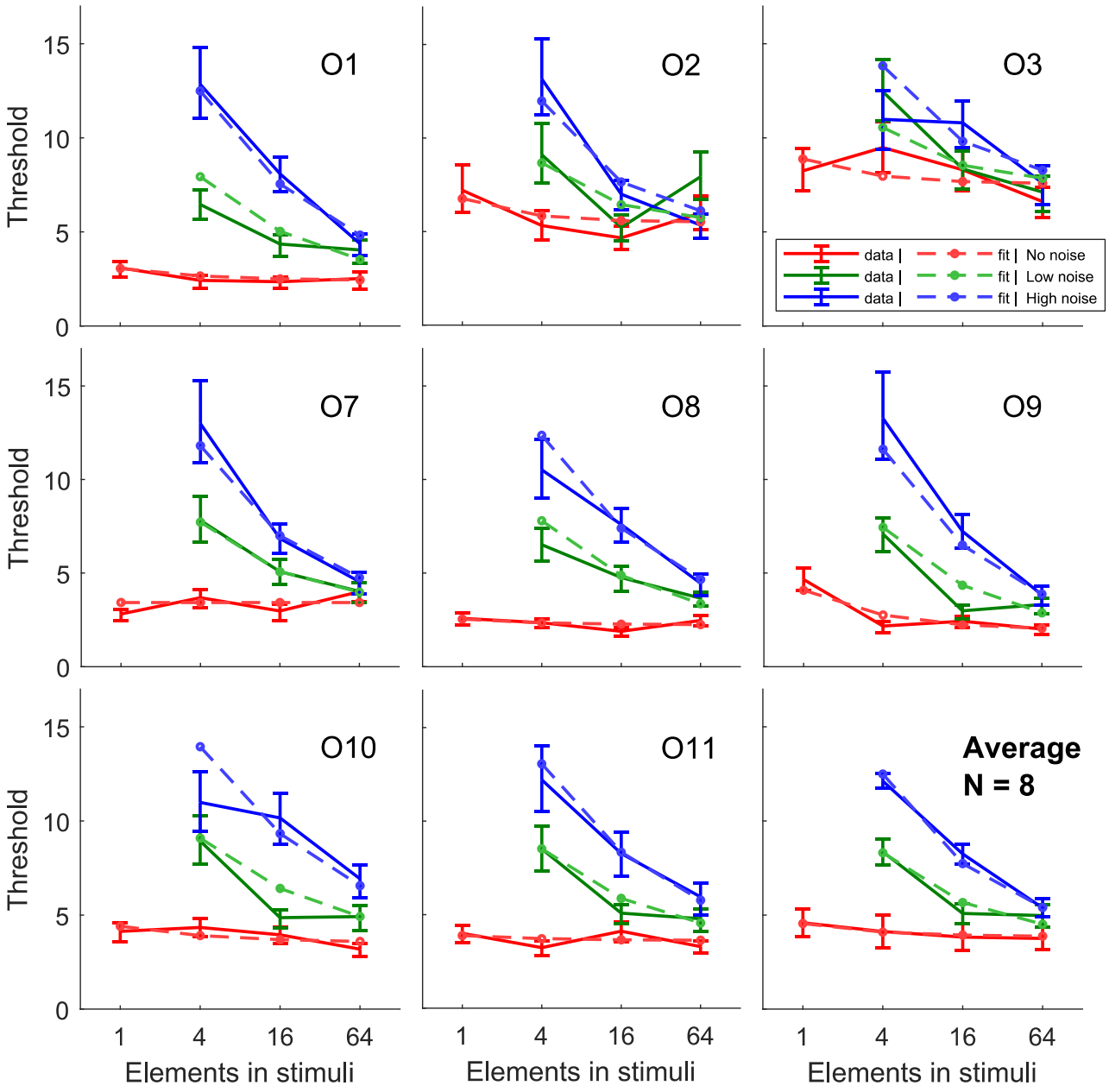

Figure 9. *Discrimination thresholds for Experiment 1. Each graph is an individual observer except the bottom right which shows observer average. X-axis indicates different set size conditions and y-axis the discrimination thresholds (PMF SDs, measured in degrees of hue angle). Different colors indicate different noise conditions. Solid lines show the discrimination thresholds from the data with error bars showing +/- 1 SEM. Dashed lines show the model predictions with best fitting parameters.*

The two variants of the equivalent noise models tested both fit the data well. The power model, however, did offer a slightly better fit for 7 out of 8 observers evaluated by log-Likelihood values. Thus, only the model fits of the power model are shown in Figure 9. Equation 2 accommodates the effects found in the experimental manipulations very well. With the power model, the power to which the number of elements is to be raised to get the observer's effective sample size varied from 0.64 to 0.89 between observers. This means that the effective sampling ranged from a minimum of 2-3 in the 4 element condition to a maximum of 16-41 in the 64-element condition between observers. The model also fit the averaged data exceptionally well, giving an estimate of 0.83 for the exponent parameter; near the higher end of the estimates for individual observers.

To explore the effects of the spatial separation of the elements, the two different 16-element conditions were compared in all the three external noise conditions. As can be seen in Figure 10, performance was roughly the same whether the elements were connected or not. The lack of effect is also reflected in the two-way repeated measures ANOVA with a nonsignificant main effect for element separation ($F(1, 7) = 1.93$, $p = .207$) and a nonsignificant interaction for element separation and external noise ($F(2, 14) = 2.08$, $p = .162$).

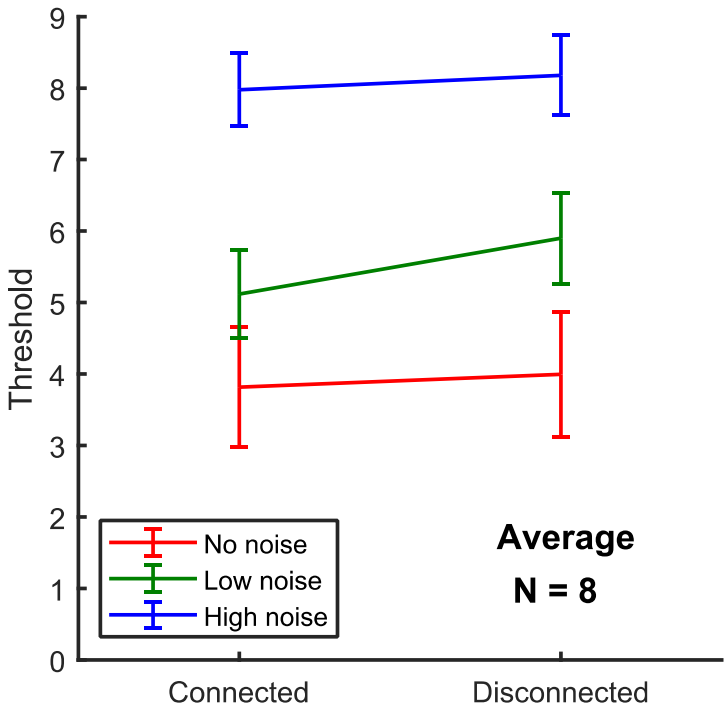

Figure 10. *Discrimination thresholds in Experiment 1 for the different 16-element conditions. On the x-axis are the spatial separation manipulations. On the y-axis are the discrimination thresholds (PMF SDs, measured in degrees of hue angle) pooled over observers and with error bars showing +/- 1 SEM. Different colors indicate different noise conditions.*

4. Experiment 2

4.1. Specific methods

As Experiment 1 confounds element number and stimulus size, in Experiment 2 we wished to ascertain that observer performance in Experiment 1 was the result of changes in element number instead of any spatial differences between the stimuli. The types of stimuli were designed to coincide in some factors while differing in others as illustrated in Figure 11. The different aspects to compare were the number of elements, signal surface size, and total stimulus size. The one element stimuli served as a control for the amount of information in a single element of different sizes. In two of the experimental conditions, the element size extended 2 degrees of visual angle instead of 1. The single element stimuli were presented with no external noise while all others had the high level of external noise.

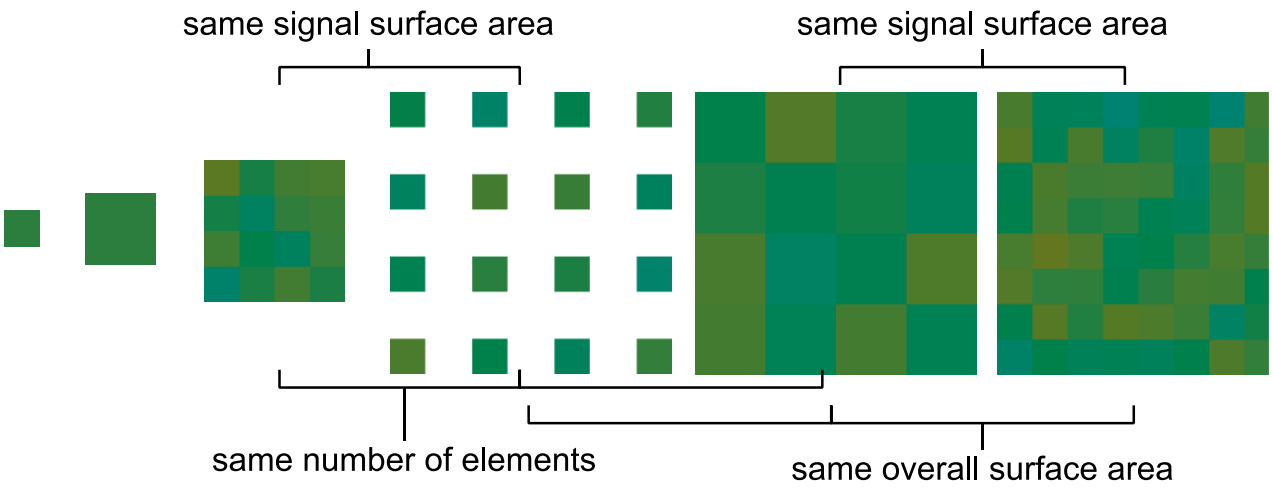

Figure 11. *Experimental conditions for Experiment 2. The experiment included six different types of stimuli. The 1-element stimuli were only tested in the no-noise condition and acted mainly as a control. The other four stimuli were designed to share some of the aspects named in the figure while differing in others. They were presented only in the high noise condition. Please note that these stimulus examples in print do not exactly match with the stimuli produced on monitor in the experiments. The experimental stimuli were also not displayed on a white background. These example stimuli are for illustration purposes only.*

Similarly to Experiment 1, Experiment 2 consisted of a practice session and the main experiment, preceded by a short demo. The practice for Experiment 2 included all six experimental conditions (see Figure 11) with 7 repetitions for each of the 9 comparison levels, resulting in 378 trials total. In the main experiment of Experiment 2, the 6 experimental blocks were repeated twice in random order, resulting in 12 total blocks. Each block had 10 repetitions for 9 comparison levels, 1080 trials in total. The whole Experiment 2 took approximately 1 hour to finish and all observers completed it in one session.

4.2. Results

The results are visualized in Figure 12. For the one element stimuli, most observers didn't have any significant difference in discrimination thresholds, giving validity to direct comparisons between stimuli with different element sizes. Of main interest was observer performance in the other four experimental conditions. Although there was some variation between observers, the only consistent effect on performance seemed to come from the number of elements in the stimulus. A three-way ANOVA (excluding the single-element conditions) with the number of elements, signal surface size and the whole stimulus size as factors, supported this result. The analysis showed a significant main effect for the number of elements ($F(1, 6) = 35.85$, $p < .001$) while all other effects were nonsignificant.

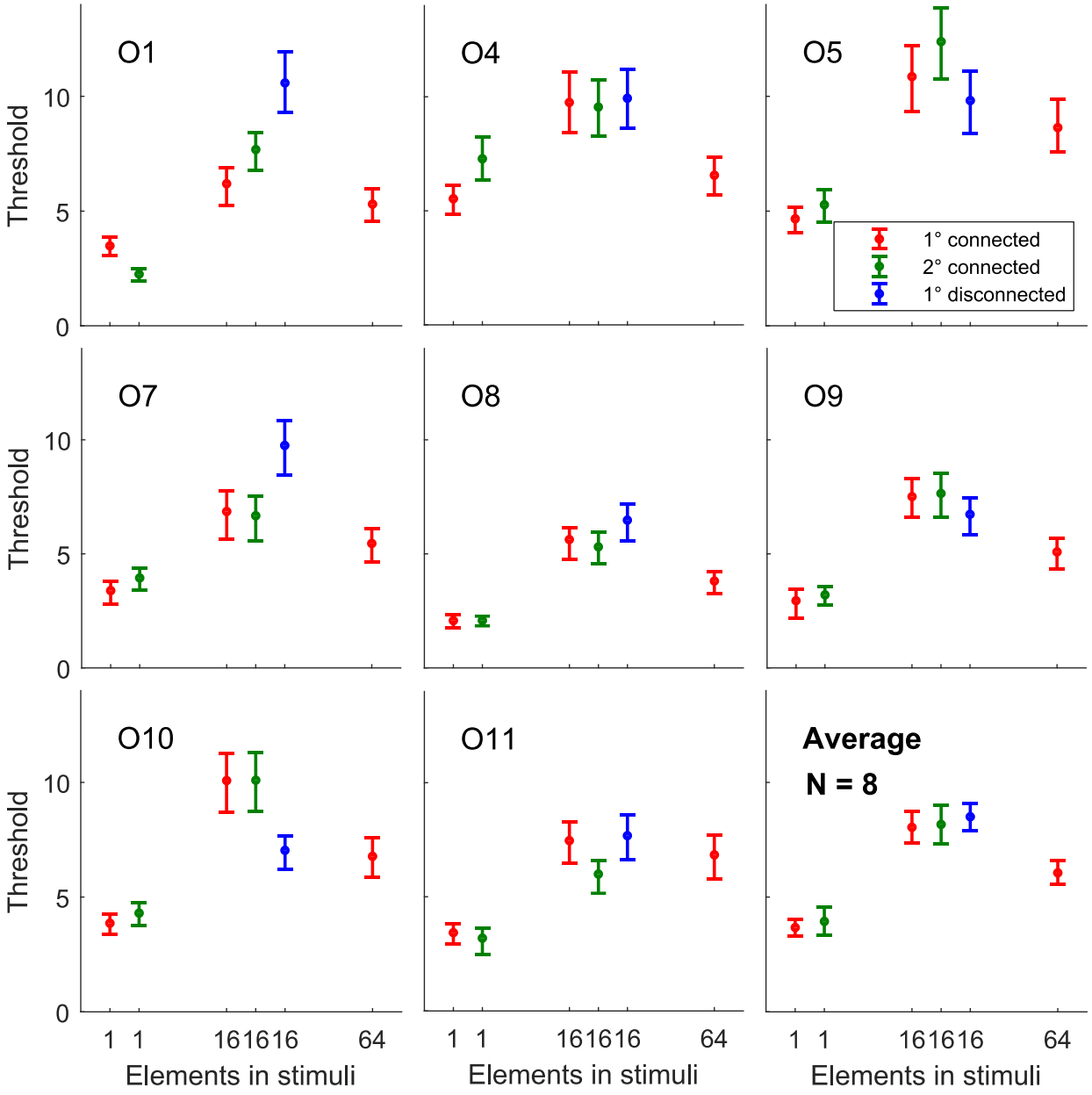

Figure 12. Discrimination thresholds in Experiment 2. Each graph is an individual observer except the bottom right which shows observer average. The different experimental conditions are shown on the x-axis while the y-axis shows the discrimination thresholds (PMF SDs, measured in hue angle). Error bars show +/- 1 SEM.

5. Experiment 3

5.1. Specific methods

For Experiment 3, the question was whether changes in the shape of the noise distribution would produce changes to observers' averaging strategy. More specifically, we wanted to find out if observers would *spontaneously* adjust their approach. For the purpose of spontaneity, we avoided any mention of "mean" or "average" in observer instructions. Instead, observers were asked to answer according to which they feel is "yellower" or "bluer", while also mentioning that there is not necessarily a correct answer.

In Experiment 3, instead of performance, the appropriate measure was the possible bias in responses. Three different set size conditions (4, 16 and 64) were selected with two levels of noise (low and high). Each of these six experimental conditions were tested both with a skewed distribution and with a baseline condition with no skew (see Figure 13). For the skewed noise distributions, element hues were drawn from a strongly skewed (skew ≈ +/- .96, switched between observers) normal distribution with a SD closely matching the von Mises κ values 40 and 15 from other experiments (SDs were approximately 9.12 and 14.79 degrees of hue angle respectively). For observers who had already completed Experiment 1, baseline measures of response bias were taken from the corresponding measures (but with connected elements). For other observers, a separate baseline measure was conducted.

In Experiment 3 the comparison stimuli were similar to Experiments 1 and 2 besides having no added external noise, meaning that all the elements were of the same hue. The reason for this was that applying the same skewed distribution to both the test and comparison stimulus would prevent disambiguating any perceptual bias, as any perceptual bias would apply equally to both stimuli. Also, using a normal distribution might have prevented observers from learning the distribution characteristics of the test stimulus. Because the comparison stimuli were of uniform hue, both the test and comparison stimuli were presented with disconnected elements (separation of 1/3 element size). This was done to avoid having observers compare stimuli with very different edge information, attached uniform elements giving the impression of a single block. In contrast, for baseline measurements the comparison stimulus hues were drawn from the same distribution as the test stimulus hues, similarly to Experiment 1.

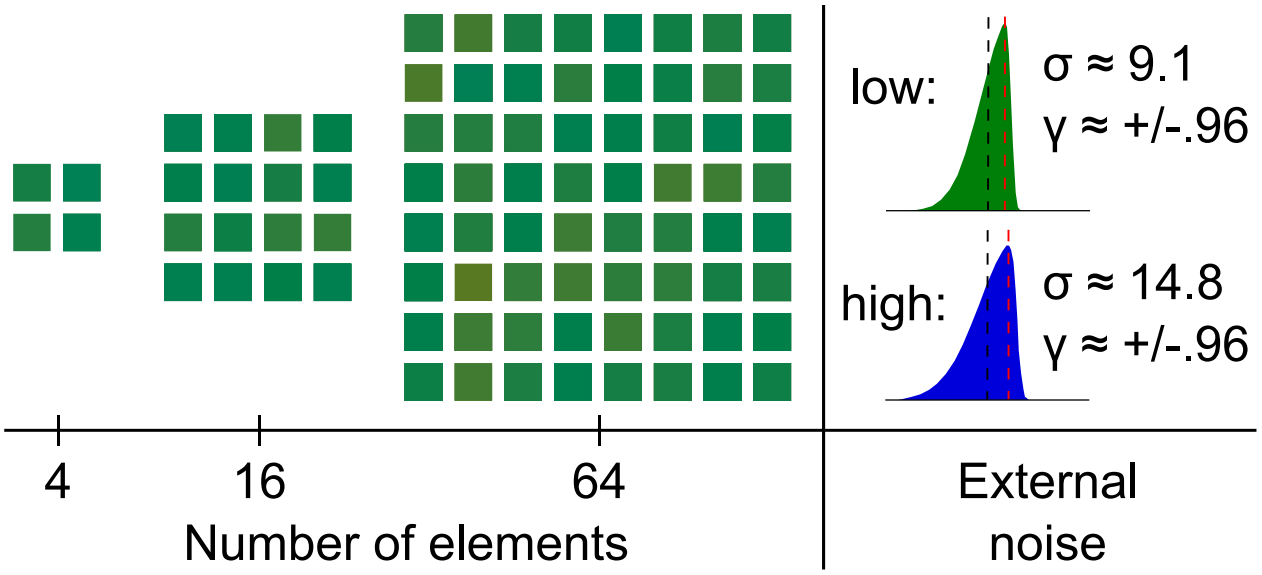

Figure 13. *Experimental conditions for Experiment 3. The experiment included three different types of stimuli as seen on the left side of the figure and two levels of added noise for the element hues on the right. The direction of the skew was switched between observers. Some observers also completed a baseline measurement which was otherwise identical, but with non-skewed noise distributions. Please note that these stimulus examples in print do not exactly match with the stimuli produced on monitor in the experiments. The experimental stimuli were also not displayed on a white background. These example stimuli are for illustration purposes only.*

Experiment 3 also consisted of a practice session and the main experiment preceded by a short demo. Practice runs had 7 repetitions for 9 comparison levels for all of the 6 experimental conditions, resulting in 378 trials total. For the main experiment of Experiment 3, there were 12 blocks in total as the 6 experiment blocks were presented in 2 intervals to gauge possible learning effects. Within an interval, the blocks were presented in a random order. Each block had 10 repetitions for 9 comparison levels, 1080 trials in total. The measures took observers approximately 1 hour. The additional baseline measures shared all the aforementioned details and similarly took observers approximately 1 hour. Furthermore, because of experimenter error, the experiment procedure only repeated the experimental blocks for the lower noise conditions four times each for the first four observers. After remedying the issue, three of the four observers returned for an additional measurement of the missing high noise condition, which included 540 trials.

5.2. Results

The results of Experiment 3 are visualized in Figure 14. Note that the results were flipped for observers who had a positively skewed distribution in their measurements. This was done to always have the skewed distribution mode in the positive direction for the sake of comparison and statistical testing. Also note that only the mode for the low noise distribution is shown in Figure 14, for the sake of visibility. Consistent biases in responses are clearly visible only for observer 10 and limited to the higher noise condition for observer 1. Most response means are located very close to the distribution mean and a two-way repeated measures ANOVA also shows no main effect for the distribution skew, with no significant deviation from zero ($F(1, 6) = 0.55$, $p = .487$). The results show very little difference from a simple averaging strategy that weighs all elements equally.

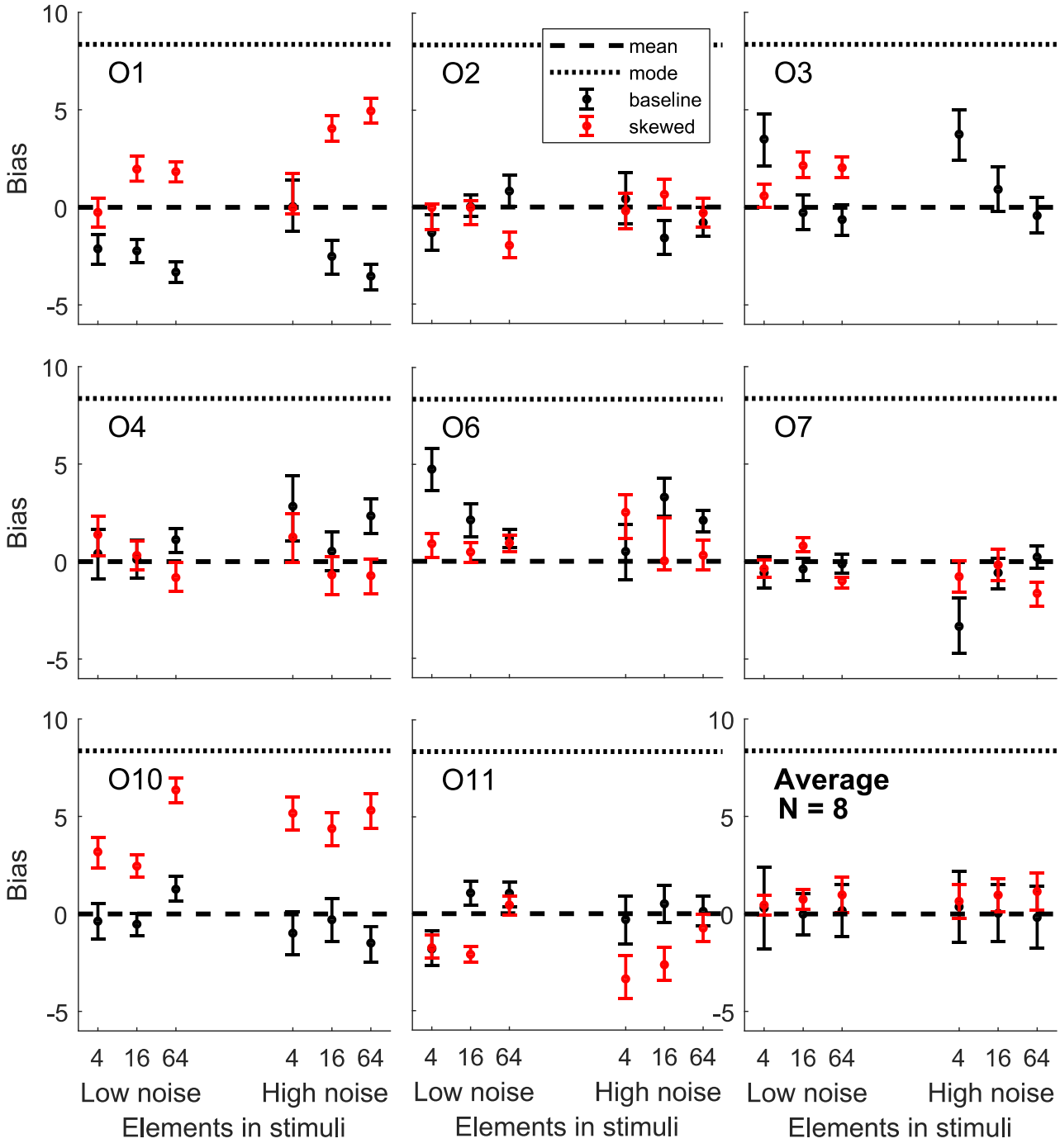

Figure 14. *Response biases in Experiment 3. Each graph is an individual observer except the bottom right which shows observer average. X-axis indicates different set size and noise conditions and y-axis the response bias (PMF means, measured in degrees of hue angle). The dashed line at the zero mark on the y-axis indicates the distribution mean and the dotted line the mode. Black dots indicate the baseline biases (with no skew) while red dots indicate biases with skewed distributions. Error bars show +/- 1 SEM. Note that although different observers were assigned with either a positively or a negatively skewed noise distribution, the results are flipped to always have the distribution mode in the positive direction.*

6. Experiment 4

6.1. Specific methods

In Experiment 4 the goal was to include the whole hue circle by employing a different type of task. Instead of the 2IFC task of previous experiments, Experiment 4 used a modified ABX comparison task. The experimental conditions also formed a more complete factorial design than Experiment 1 in which the number of elements (1, 4, 16, 36 and 64), spatial separation between elements (connected, disconnected by a distance of 1/3 element size), and amount of external noise in the hue of elements (no noise, low noise, high noise) served as independent variables. Following the factorial design, stimulus blocks were formed according to the unique permutations of the factorial variables (see Figure 15). Excluding 3 redundant stimulus blocks, this came to a total of 27 blocks.

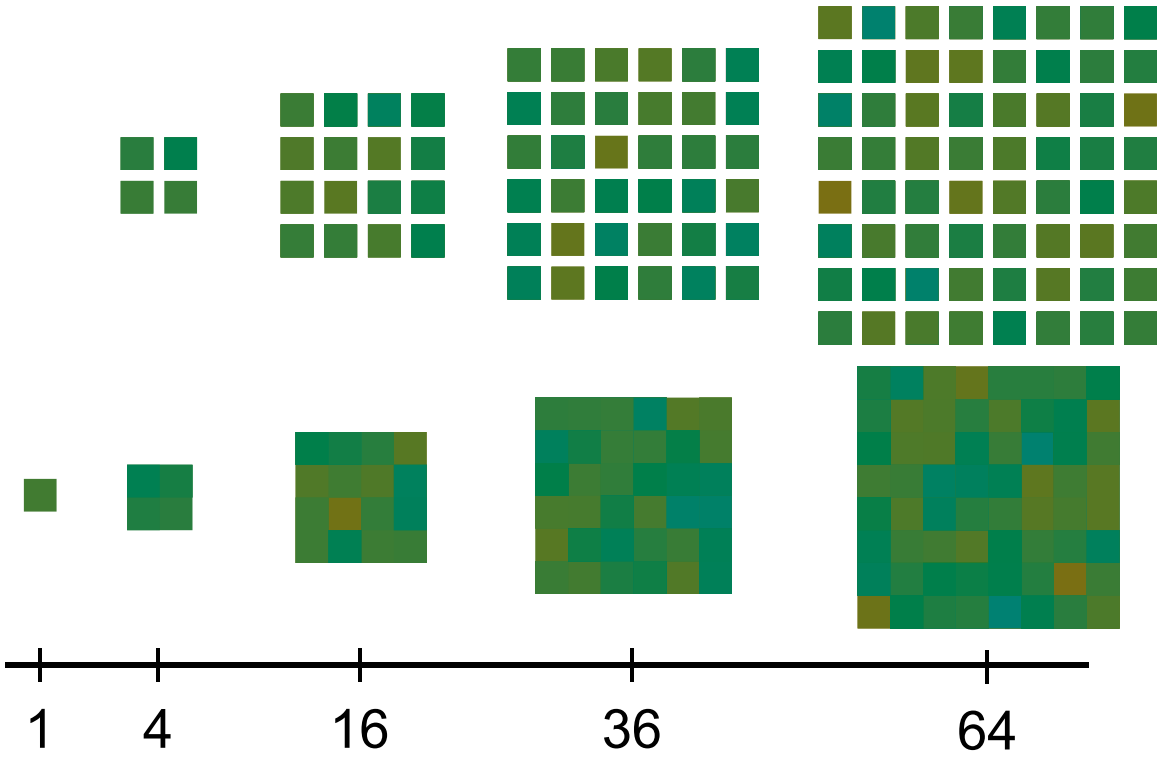

Figure 15. *Experimental conditions for Experiment 4. The experiment included nine different types of stimuli and three levels of added noise for the element hues (similarly to Experiment 1), in a factorial design. Please note that these stimulus examples in print do not exactly match with the stimuli produced on monitor in the experiments. The experimental stimuli were also not displayed on a white background. These example stimuli are for illustration purposes only.*

Instead of one comparison stimulus, Experiment 4, had two comparison stimuli filled with a uniform hue. The squares extended 4 degrees of visual angle, were centered to the screen vertically and were separated horizontally by a distance of 2 degrees of visual angle to either side of the screen center. There were 9 comparison levels around the test stimulus mean, and the comparison stimuli were separated from the comparison levels by an equal hue angle into opposite directions (see Figure 16). The values for the comparison levels and the separation of the comparison stimuli were determined by the observer's absolute discrimination threshold. For the practice sessions and the main experiment of Experiment 4, the observer's task was to indicate which of the two comparison stimuli of uniform hue was more similar to the hue represented by the whole of test stimulus elements. Each trial was similar to Experiments 1, 2 and 3, but the center dot and inter-stimulus intervals lasted for 300 ms and stimulus presentations for 800 ms.

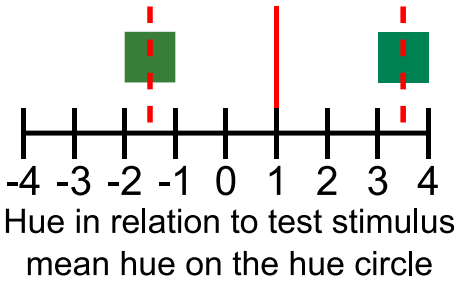

Figure 16. *Comparison stimuli in Experiment 4. The x-axis indicates the difference in hue from the test stimulus mean as comparison levels. For example, in the case of comparison level +1 indicated by the solid red line, the two comparison stimuli would have their hues from an equal difference in opposite directions on the hue circle indicated by the dashed red lines. The exact difference between the comparison stimuli hues differed between observers.*

For Experiment 4, there was also an initial odd-one-out discrimination threshold task. The stimulus consisted of 3 square uniform color fields. The squares extended 4 degrees of visual angle and were set in a triangle formation around the middle of the screen. Two of the fields were filled with the same hue and one filled with a hue offset by 1-9 (comparison levels) times 0.02 or - 0.02 radians (approximately +/- 1.15°) of hue angle. The observer was instructed to indicate which of the three stimuli differed from the other two. In each trial the observer was shown a blank screen (uniform gray of the background) with a white dot of 0.1 degrees of visual angle size indicating the middle of the screen for 300 ms. Next, the 3 stimuli were displayed for 800 ms after which the screen was blank until the observer gave a response. There were 21 repetitions for each of the 9 comparison levels in a random order resulting in 189 trials total.

Experiment 4 consisted of three parts: color discrimination measurements, practice sessions, and the main experiment. For the first experiment session, the observer started with a threshold measurement, followed by a practice set, another threshold measurement and another practice set. For following sessions, practice and threshold measurements were only performed once in the beginning. In the main experiment of Experiment 4, each block was repeated twice in a random order, resulting in 54 experiment blocks. For practice sessions, each comparison level had 3 repetitions for the 6 experimental conditions included, resulting in 162 trials total. For the main experiment, a trial block contained 10 repetitions for each of the 9 comparison levels, 4860 trials in total. All the measurements in Experiment 4 took observers approximately 4 hours in total, and were conducted in 2-4 separate sessions.

6.2. Results

Despite the change in task, the same observer who found Experiment 1 difficult, also found Experiment 4 difficult to perform. This observer's discrimination thresholds indicated a sharp decline in performance with increasing set size regardless of external noise level. There were several experimental conditions where the two different measures of the same experimental block had largely different quality of fits and/or resulted in very different estimates. Again, the data for this particular observer was omitted from the analysis. For another observer there was one measurement of an experimental block with a failure to save data. For this particular case, only the other measurement was used for fitting a PMF.

The results for the remaining five observers are visualized in Figure 17. There was high variability within observers and performance remained very similar with increasing set sizes, even with the high noise level. Despite this, a three-way repeated measures ANOVA showed significant main effects for external noise ($F(2, 8) = 104.30$, $p < .001$) and set size ($F(3, 12) = 10.23$, $p = .001$), but not their interaction ($F(6, 24) = 2.29$, $p = .068$). Also, notably, the main effect of the spatial separation of elements did not reach statistical significance ($F(1, 4) = 2.01$, $p = .229$). In comparison to Experiment 1, the very clear and consistent effects of increasing the number of elements are less obvious from visual inspection.

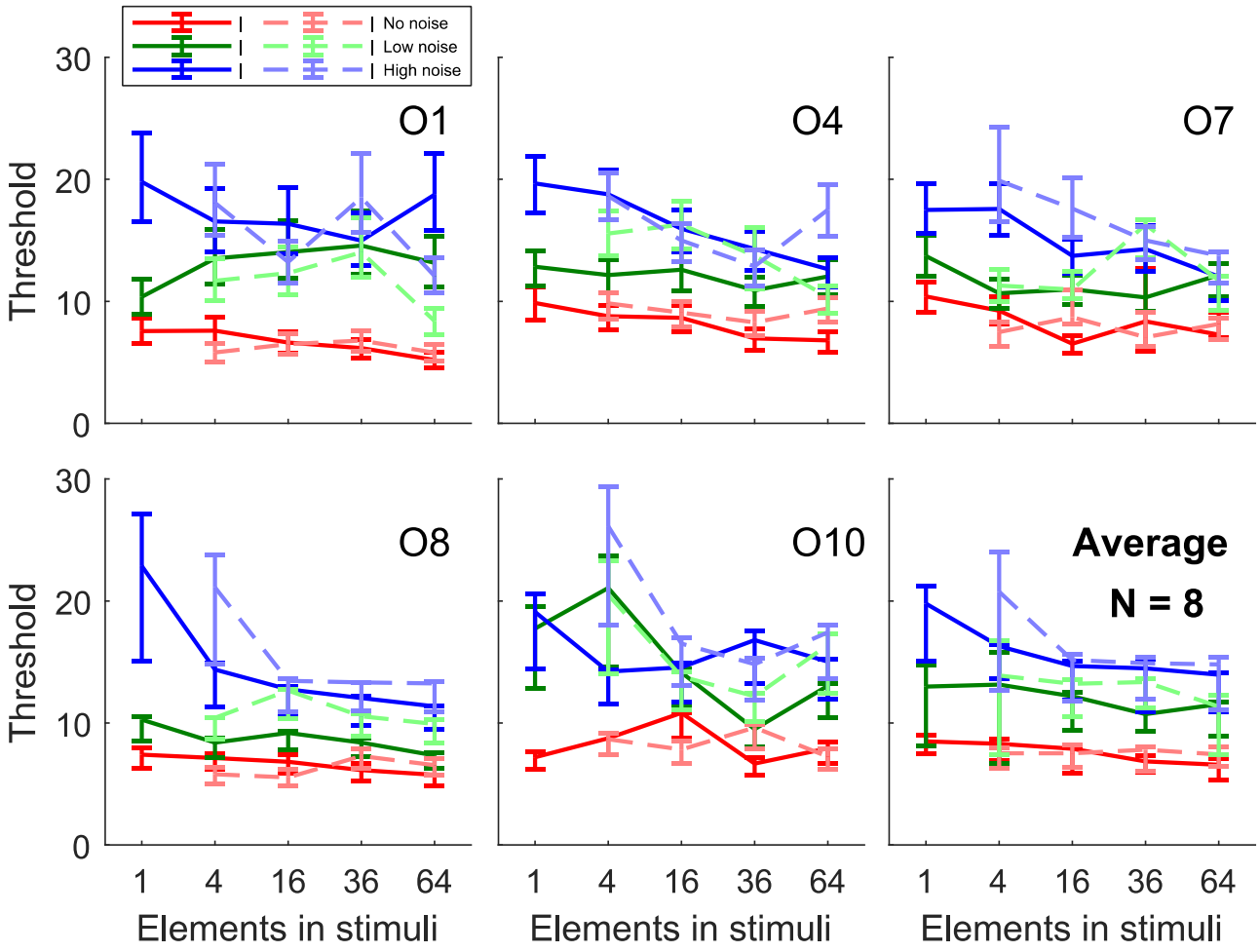

Figure 17. *Discrimination thresholds for Experiment 4. Each graph is an individual observer except the bottom right which shows observer average. X-axis indicates different set size conditions and y-axis the discrimination thresholds (PMF SDs, measured in degrees of hue angle). Different colors indicate different noise conditions. Solid lines indicate conditions with connected elements, while dashed lines indicate the disconnected conditions. Error bars show +/- 1 SEM.*

7. Discussion

In summary, our results in Experiment 1 showed that observers were effective at averaging hue information when the amount of elements increased. This effectiveness was more clearly visible with higher external noise (more uncertainty in each element). In Experiment 2, we confirmed that the improved performance was driven by the number of elements, not stimulus size. In tandem, we discovered that none of the spatial manipulations in our experiments had any noticable effect to the averaging process. In Experiment 3, we found that most observers spontaneously opted for simple

averaging with equal weights, despite the fact that the distribution was strongly skewed. Surprisingly, in Experiment 4, we found that a change in the observers' task impaired observers' effectiveness in hue averaging considerably. In the less straightforward comparison, observers' performance was also much less consistent.

It has been previously suggested that hue averaging can be accounted for by a limited random subsampling mechanism using just two elements for averaging (Maule & Franklin, 2016). Our results from Experiment 1 were in sharp contrast to such a mechanism. Not only did averaging performance increase from 16-element stimuli to 64-element stimuli, our modeling results pointed to a maximum effective sampling size somewhere between 16 and 41 elements. Most of the difference in results between these two studies is probably due to the difference in stimuli. Maule and Franklin (2016) employed a limited set (maximum 4 different hues) of clearly distinguishable hues, which limited the relevant info for the observer and could also encourage a more cognitive strategy. Tokita et al. (2016) proposed that observers have large differences in strategy when estimating summary statistics from small sets and very similar strategies with larger sets. The different task in our Experiment 1 and Maule and Franklin's (2016) study could also play a major role. As we found out in Experiment 4, switching from a very straightforward comparison task to a slightly indirect comparison hindered averaging performance.

Our modeling results showed that the observer data from Experiment 1 fit the equivalent noise framework very well. A slightly better fit was achieved when the amount of samples observers utilized was defined as a proportion of available elements instead of a fixed maximum, similarly to Dakin (2001) and also suggested by Whitney and Yamanashi Leib (2018). Based on either version of the model, the number of elements observers utilized clearly surpassed the limits of attentional (Scimeca & Franconeri, 2015) and working memory resources (Luck & Vogel, 2013) assuming each element was attended serially. Therefore, the results support a more global mechanism with distributed attention in averaging hue, which has been suggested in one way or another in most other domains of ensemble perception (e.g. Alvarez & Oliva, 2008; Chong & Treisman, 2003; Corbett & Oriet, 2011; Im & Halberda, 2013). Another possibility, that was not directly assessed in this study, is using a heuristic for the decision criterion that could reduce the number of necessary samples enough to be handled by directed attention (see Lau & Brady, 2018). However, it is unknown if a useful heuristic exists for the task in Experiment 1 and it would require smart subsampling (directing sampling to the most meaningful targets). Without clear extremes or outliers that "pop out", smart subsampling could be of limited effectiveness.

Sampling is a highly relevant question also when considering the results from Experiment 2. What we found was that the improvement in hue averaging was only driven by the number of hue elements. None of the spatial manipulations seemed to consistently affect hue averaging. This is by no means a trivial result, as the different types of stimuli had large differences in information conveyed. One possible interpretation is that in each case the hue elements were cognitively processed as individual pieces of information, therefore making them equal despite the differences in spatial signal. However, our modeling results did not support serial processing. Another possible interpretation is that the hue ensembles were processed (more or less) globally in a way that ignored spatial relations. This could easily be related to texture-like processing as suggested by Im and Halberda (2013). Using an inhibition release paradigm in a functional magnetic resonance imaging (fMRI) study, Cant & Xu (2014) experimented with different kinds of ensemble-like images. Opposing texture-like processing, the areas of the brain sensitive to changes in summary statistics of ensembles and textures were not sensitive to changes in color ratio. Nevertheless, as Cant and Xu only probed the question in a very particular setting with clearly identifiable high-level stimuli, it is unclear whether the responses to color changes were simply too weak to identify in fMRI.

Whatever the exact mechanism might be, global sampling still leaves us with two distinct possibilities. Either hue distributions are perceived similarly for scenes and objects, or our stimuli with connected elements were not considered unified objects. In our experiments, there was no illumination or 3-dimensional cues to individuate objects, just spatial edge contrast. Would such edge contrast be ignored in ensemble perception of hue? Ignoring edge contrasts could indeed support swift image segmentation in a way suggested by Utochkin (2015). As such, ensemble perception of color could act as a parallel or a joint mechanism in fast scene perception. Further study is required to find out whether any change in hue integration occurs when the stimulus is clearly indicated as a single object. Also, definitive conclusions in sampling could be probed with individual elements that by themselves are not clearly perceived.

It has been previously suggested by Chetverikov et al. (2017) that observers are able to form representations of a color distribution shape. Their results are derived from observers' reaction times in an outlier search task, in which the learned representation of a distractor distribution would aid (or hinder) the search. However, it is especially noteworthy that observers received feedback on their search result (as well as search time). The swift improvement during the first few trials with a similar distribution could be either of a detailed representation of the distractor distribution or sharpening of the task and decision criteria involved, or of course, a little bit of both. The different profiles of reaction times after learning either a Gaussian or a uniform distractor distribution, could

similarly reflect a more clear-cut (and wide) dismissal range by a learned decision criterion. Whether this requires a detailed representation of the underlying distribution or not, would appear yet inconclusive.

In contrast to learning, we set out to examine if observers would spontaneously form skewed representations (indirectly, is the ensemble percept rich enough to support such inferences). Only two of the observers showed any sign of adjusting their strategy from simple averaging. In these two cases the most bias was seen with higher noise and larger set sizes, which would have included more information about the shape of the distribution while also having a larger possibility for significant outliers. Arguably, as the found biases were still far from the distribution modes, the results could be consolidated by some observers simply discarding the most extreme outliers. Considering that six of the eight observers were effectively unbiased, it is not feasible to say that even discarding outliers would be an obvious spontaneous strategy for observers. Much more consistent discarding of outliers has been reported with facial expression (Haberman & Whitney, 2010). The difference may lie in the extremity of the outliers. With the facial expressions, outliers were clearly displaying a whole different emotion. This could be somewhat analogous to crossing color category borders. Would observers be more likely to discard outliers if they crossed color categories? This could be seen as a more direct violation of an inclusion criterion assuming statistical image segmentation (Utochkin, 2015). Or is it possible that instead of depending on color categories, there exists a particular decision criterion? If so, could this be flexibly adjusted to aid the visual task in hand?

Besides all the previous considerations, the results of Experiment 4 add to the complexities of hue averaging. In Experiment 4 we found generally similar results as in Experiment 1 with increasing number of elements improving hue averaging. However, there was much more random variation in the obtained discrimination thresholds, and the clear effects seen in Experiment 1 were not nearly as consistent in Experiment 4. The hue averaging part of the observers' task was nearly identical to Experiment 1, so what causes such a distinct difference in the results? There are some readily available hypotheses. First of all, there could be difficulty in "splitting" the comparison between two comparison stimuli. Besides the obvious toll of dividing one's attention, many trials had observers compare the comparison stimuli in separate directions of hue angle compared to the test stimulus. The previous is also connected to the second point, that observers might find it difficult to estimate "distances" in hue space. Instead of straightforward comparisons, in Experiment 4 observers were forced to sometimes estimate an arbitrary distance in an abstract space that does not have a general real-life metric. A third issue worth consideration is that a task of forming ensemble

percepts (especially large and variable) could prime the visual system into distributed spatial processing. As spatial properties did not guarantee segmentation of objects in Experiment 2, the comparison stimuli in Experiment 4 could have suffered from partial blending of hues. It is unclear whether any of these issues could have caused the difference in results between Experiment 1 and Experiment 4, especially when they all relate to the task of comparison, while the issue is mainly of effectiveness of averaging.

Overall, observers are effective at averaging hue distributions and this process seems to ignore many spatial aspects of the stimulus. The hue of an object with varying surface reflectance is therefore not determined from singular sample locations, but averaged over larger areas, even if the percept is not spatially continuous. Simple averaging emerges as the most common spontaneous estimation strategy even with a non-normal hue distribution. The details of the observer's task are also far from trivial when we wish to estimate the effectiveness of hue averaging. A more straightforward task comparing hue distributions suggests more effective averaging than comparing a hue distribution to uniform hues. It seems that hue averaging shares many characteristics of ensemble perception overall, but this does not generalize in a straightforward manner to all situations. As such, in future studies it should be carefully considered how even the minor task details might affect ensemble perception in any given domain. Color also presents unique difficulties and complexities for being multidimensional, relative and highly adaptable. But with complexities, there are also venues to further our understanding on how exactly the perception of color aids and guides us in different ways in different situations.

References

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257-262.

Alvarez, G. A. (2011). Representing multiple objects as an ensemble enhances visual cognition. *Trends in cognitive sciences*, *15*(3), 122-131.

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological science*, *19*(4), 392-398.

Ariely, D. (2001). Seeing sets: Representation by statistical properties. *Psychological science, 12*(2), 157-162.

Balaraman, S. (1962). Color vision research and trichromatic theory: A historical review. *Psychological bulletin, 59*(5), 434.

Bauer, B. (2015). A selective summary of visual averaging research and issues up to 2000. *Journal of vision, 15*(4), 14-14.

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance, 43*(6), 1160.

Brainard, D. H. (1989). Calibration of a computer controlled color monitor. *Color Research & Application, 14*(1), 23-34.

Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision, 10*, 433-436.

Brainard, D. H., Cottaris, N. P., & Radonjić, A. (2018). The perception of colour and material in naturalistic tasks. *Interface Focus, 8*(4), 20180012.

Brainard, D. H., & Stockman, A. (2010). *Colorimetry*. McGraw Hill.

Broadbent, A. D. (2004). A critical review of the development of the CIE1931 RGB color-matching functions. *Color Research & Application, 29*(4), 267-272.

Cant, J. S., Large, M. E., McCall, L., & Goodale, M. A. (2008). Independent processing of form, colour, and texture in object perception. *Perception*, *37*(1), 57-78.

Cant, J. S., & Xu, Y. (2012). Object ensemble processing in human anterior-medial ventral visual cortex. *Journal of Neuroscience*, *32*(22), 7685-7700.

Cant, J. S., & Xu, Y. (2014). The impact of density and ratio on object-ensemble representation in human anterior-medial ventral visual cortex. *Cerebral Cortex*, *25*(11), 4226-4239.

Cant, J. S., & Xu, Y. (2017). The contribution of object shape and surface properties to object ensemble representation in anterior-medial ventral visual cortex. *Journal of cognitive neuroscience*, *29*(2), 398-412.

Cavina-Pratesi, C., Kentridge, R. W., Heywood, C. A., & Milner, A. D. (2010). Separate channels for processing form, texture, and color: evidence from fMRI adaptation and visual object agnosia. *Cerebral cortex*, *20*(10), 2319-2332.

Chetverikov, A., Campana, G., & Kristjánsson, Á. (2017b). Representing color ensembles. *Psychological science*, *28*(10), 1510-1517.

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision research*, *43*(4), 393-404.

Chong, S. C., & Treisman, A. (2005). Attentional spread in the statistical processing of visual displays. *Perception & Psychophysics*, *67*(1), 1-13.

Corbett, J. E., & Melcher, D. (2014). Stable statistical representations facilitate visual search. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(5), 1915.

Corbett, J. E., & Oriet, C. (2011). The whole is indeed more than the sum of its parts: Perceptual averaging in the absence of individual item representation. *Acta psychologica*, *138*(2), 289-301.

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why?. *Current directions in psychological science*, *19*(1), 51-57.

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic bulletin & review*, *24*(4), 1158-1170.

Dakin, S. C. (2001). Information limit on the spatial integration of local orientation signals. *JOSA A*, *18*(5), 1016-1026.

Dakin, S., & Wagemans, J. (2015). Seeing statistical regularities: Texture and pattern perception. *The oxford handbook of perceptual organization*, 150-167.

de Fockert, J., & Wolfenstein, C. (2009). Rapid extraction of mean identity from sets of faces. *The Quarterly Journal of Experimental Psychology*, *62*(9), 1716-1722.

Derrington, A. M., Krauskopf, J., & Lennie, P. (1984). Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology*, *357*(1), 241-265.

Ernst, M. O., & Banks, M. S. (2002). Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, *415*(6870), 429.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598.

Fairman, H. S., Brill, M. H., & Hemmendinger, H. (1997). How the CIE 1931 color-matching functions were derived from Wright-Guild data. *Color Research & Application, 22*(1), 11-23.

Foster, D. H. (2011). Color constancy. *Vision research*, *51*(7), 674-700.

Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual review of neuroscience*, *26*(1), 181-206.

Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, *17*(17), R751-R753.

Haberman, J., & Whitney, D. (2009). Seeing the mean: ensemble coding for sets of faces. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(3), 718.

Haberman, J., & Whitney, D. (2010). The visual system discounts emotional deviants when extracting average expression. *Attention, Perception, & Psychophysics*, *72*(7), 1825-1838.

Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature neuroscience*, *9*(11), 1367.

Hillis, J. M., Ernst, M. O., Banks, M. S., & Landy, M. S. (2002). Combining sensory information: mandatory fusion within, but not between, senses. *Science*, *298*(5598), 1627-1630.

Hillis, J. M., Watt, S. J., Landy, M. S., & Banks, M. S. (2004). Slant from texture and disparity cues: Optimal cue combination. *Journal of vision*, *4*(12), 1-1.

Hurvich, L. M., & Jameson, D. (1957). An opponent-process theory of color vision. *Psychological review*, *64*(6p1), 384.

Im, H. Y., & Halberda, J. (2013). The effects of sampling and internal noise on the representation of ensemble average size. *Attention, Perception, & Psychophysics*, *75*(2), 278-286.

Ishihara, S. (1973). Test for Colour-Blindness, 24 Plates Edition, Kanehara Shuppan Co. *Ltd., Tokyo*.

Kimura, E. (2018). Averaging colors of multicolor mosaics. *JOSA A*, *35*(4), B43-B54.

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, *36*(14), 1.

Knill, D. C., & Saunders, J. A. (2003). Do humans optimally integrate stereo and texture information for judgments of surface slant?. *Vision research*, *43*(24), 2539-2558.

Kuriki, I. (2004). Testing the possibility of average-color perception from multi-colored patterns. *Optical review*, *11*(4), 249-257.

Landy, M. S., & Kojima, H. (2001). Ideal cue combination for localizing texture-defined edges. *JOSA A*, *18*(9), 2307-2320.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision research*, *35*(3), 389-412.

Lau, J. S. H., & Brady, T. F. (2018). Ensemble statistics accessed through proxies: Range heuristic and dependence on low-level properties in variability discrimination. *Journal of vision*, *18*(9), 3-3.

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in cognitive sciences*, *17*(8), 391-400.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347.

Maule, J., & Franklin, A. (2015). Effects of ensemble complexity and perceptual similarity on rapid averaging of hue. *Journal of vision*, *15*(4), 6-6.

Maule, J., & Franklin, A. (2016). Accurate rapid averaging of multihue ensembles is due to a limited capacity subsampling mechanism. *JOSA A*, *33*(3), A22-A29.

Maule, J., Witzel, C., & Franklin, A. (2014). Getting the gist of multiple hues: metric and categorical effects on ensemble perception of hue. *JOSA A*, *31*(4), A93-A102.

Milojevic, Z., Ennis, R., Toscani, M., & Gegenfurtner, K. R. (2018). Categorizing natural color distributions. *Vision research*.

Mollon, J. D., & Jordan, G. (1997). On the nature of unique hues. *John Dalton's colour vision legacy*, 381-392.

Myczek, K., & Simons, D. J. (2008). Better than average: Alternatives to statistical summary representations for rapid judgments of average size. *Perception & psychophysics*, *70*(5), 772-788.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, *155*, 23-36.

Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of vision*, *8*(5), 13-13.

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature neuroscience*, *4*(7), 739.

Rentzeperis, I., Nikolaev, A. R., Kiper, D. C., & van Leeuwen, C. (2014). Distributed processing of color and form in the visual cortex. *Frontiers in psychology*, *5*, 932.

Robertson, A. R. (1977). The CIE 1976 color-difference formulae. *Color Research & Application*, *2*(1), 7-11.

Saarela, T. P., & Landy, M. S. (2012). Combination of texture and color cues in visual segmentation. *Vision research*, *58*, 59-67.

Saarela, T. P., & Landy, M. S. (2015). Integration trumps selection in object recognition. *Current Biology*, *25*(7), 920-927.

Scimeca, J. M., & Franconeri, S. L. (2015). Selecting and tracking multiple objects. *Wiley Interdisciplinary Reviews: Cognitive Science*, *6*(2), 109-118.

Shapley, R., & Hawken, M. J. (2011). Color in the cortex: single-and double-opponent cells. *Vision research*, *51*(7), 701-717.

Smith, V. C., & Pokorny, J. (2003). Color matching and color discrimination. *The science of color*, *2*, 103-148.

Stockman, A., & Brainard, D. H. (2010). Color vision mechanisms. *OSA handbook of optics*, 11-11.

Sunaga, S., & Yamashita, Y. (2007). Global color impressions of multicolored textured patterns with equal unique hue elements. *Color Research & Application*, *32*(4), 267-277.

Tokita, M., Ueda, S., & Ishiguchi, A. (2016). Evidence for a global sampling process in extraction of summary statistics of item sizes in a set. *Frontiers in psychology*, *7*, 711.

Utochkin, I. S. (2015). Ensemble summary statistics as a basis for rapid visual categorization. *Journal of Vision*, *15*(4), 8-8.

Utochkin, I. S., & Vostrikov, K. O. (2017). The numerosity and mean size of multiple objects are perceived independently and in parallel. *PloS one*, *12*(9), e0185452.

Victor, J. D., Conte, M. M., & Chubb, C. F. (2017). Textures as probes of visual processing. *Annual review of vision science*, *3*, 275-296.

Watamaniuk, S. N., & McKee, S. P. (1998). Simultaneous encoding of direction at a local and global scale. *Perception & Psychophysics*, *60*(2), 191-200.

Watamaniuk, S. N., Sekuler, R., & Williams, D. W. (1989). Direction perception in complex dynamic displays: the integration of direction information. *Vision research*, *29*(1), 47-59.

Webster, J., Kay, P., & Webster, M. A. (2014). Perceiving the average hue of color arrays. *JOSA A*, *31*(4), A283-A292.

Whitney, D., & Leib, A. Y. (2018). Ensemble perception. *Annual review of psychology*, *69*.

Wichmann, F. A., & Hill, N. J. (2001). The psychometric function: II. Bootstrap-based confidence intervals and sampling. *Perception & psychophysics*, *63*(8), 1314-1329.

Witzel, C., & Gegenfurtner, K. R. (2015). Categorical facilitation with equally discriminable colors. *Journal of vision*, *15*(8), 22-22.

Witzel, C., & Gegenfurtner, K. R. (2016). Categorical perception for red and brown. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(4), 540.

Witzel, C., & Gegenfurtner, K. R. (2018). Color Perception: Objects, Constancy, and Categories. *Annual review of vision science*, *4*, 475-499.

Wyszecki, G., & Stiles, W. S. (1982). *Color science (Vol. 8).* New York: Wiley.

Yang, Y., Tokita, M., & Ishiguchi, A. (2018). Is There a Common Summary Statistical Process for Representing the Mean and Variance? A Study Using Illustrations of Familiar Items. *i-Perception*, *9*(1), 2041669517747297.

Young, M. J., Landy, M. S., & Maloney, L. T. (1993). A perturbation analysis of depth perception from combinations of texture and motion cues. *Vision research*, *33*(18), 2685-2696.